PONTIFICIAE ACADEMIAE SCIENTIARVM SCRIPTA VARIA
98

SCIENCE FOR SURVIVAL AND SUSTAINABLE DEVELOPMENT

the

PROCEEDINGS

of

the Study-Week of the Pontifical Academy of Sciences
12-16 March 1999

PONTIFICIA
ACADEMIA
SCIENTIARVM

EX AEDIBVS ACADEMICIS IN CIVITATE VATICANA

MM

PONTIFICIAE ACADEMIAE SCIENTIARVM SCRIPTA VARIA
98

SCIENCE FOR SURVIVAL AND SUSTAINABLE DEVELOPMENT

the
PROCEEDINGS
of

the Study-Week of the Pontifical Academy of Sciences
12-16 March 1999

PONTIFICIA
ACADEMIA
SCIENTIARVM

EX AEDIBVS ACADEMICIS IN CIVITATE VATICANA

MM

Casina Pio IV - Vatican Gardens
Headquarters of the Pontifical Academy of Sciences

The opinions freely expressed during the presentation of papers in the study-week, although published by the Pontifical Academy of Sciences, only represent the points of view of the participants and not those of the Academy.

Editors of the Proceedings:

V.I. KEILIS-BOROK and M. SÁNCHEZ SORONDO

ISBN 88-7761-071-9

CONTENTS

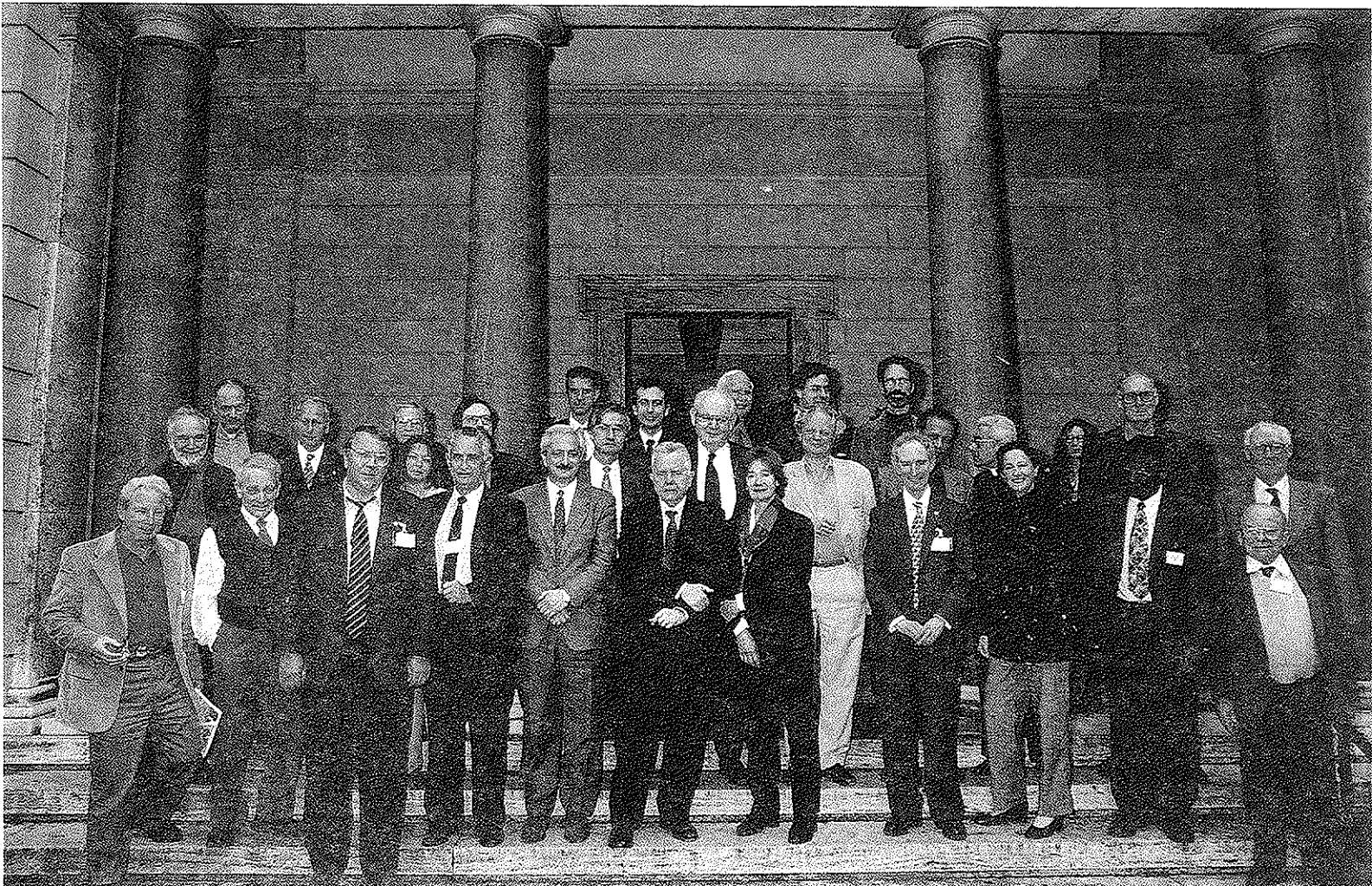
SCIENTIFIC PAPERS

I.
Problems of Sustainability: Response to the Threats of the Time Scale of Decades

II.
Problems of Mankind's Survival: Response to the Threat of Catastrophes which can Happen at any Moment

III.
Scenarios of Transitions to Critical Phenomena

IV.
Sciences and Public Policy



INTRODUCTION

"Men and women have at their disposal an array of resources for generating greater knowledge of truth so that their lives may be even more human" (*Fides et Ratio*, 'Know Yourself', § 5). In this spirit the Pontifical Academy of Sciences organized a study-week on the subject of *science for survival and sustainable development* from 12-16 March of this year. In addressing itself to how to use the many resources of modern knowledge, the Academy paid especial attention to those rooted in non-linear dynamics – the science of chaos and self-organization. This branch of learning now finds itself on the very frontiers of basic research. It is a discipline which studies chaotic systems of interacting elements (entities) which overcome their own complexity and organize themselves into strange, counter-intuitive but clearly recognizable behaviour patterns, much like those to be found in the fascinating configurations made by a waterfall or a log fire. These patterns sometimes culminate in abrupt overall changes which may be termed "critical phenomena".

The science of chaos and self-organization originated about a hundred years ago. Like other forms of basic research, it was driven forward in the first instance by the human thirst for pure knowledge. However, sharing the common destiny of basic research, this branch of science eventually acquired a paramount practical importance. This is because a critical phenomenon in some chaotic systems may mean *furia degli elementi* and industrial and socio-economic disasters.

Without always being aware of the fact, it can be said that we are surrounded by chaotic systems: the earth's crust with its millions of billions of billions of grains of rock which self-organize from time to time into a devastating earthquake; a megalopolis on its way to self-destruction; a socio-economic system prone to an outburst of mass violence or economic collapse, etc. Our world becomes more and more vulnerable to such disasters which are always on the horizon and still take us by surprise. They may take place at any moment, even while you are reading this introduction: causing up to a million casualties, rendering a large part of our world uninhabitable, triggering global economic depression, or sparking off a war in a "hot"

region. As threats to the survival of our civilization, such disasters are commonly placed on the same level as nuclear war.

Of equal importance is the question of the sustainability of our world over the next decades. Our planet is threatened by a multitude of interacting processes – the depletion of natural resources; climatic changes; population growth (from 2.5 billion people to over 6 billion over the last 50 years); a rapidly growing disparity in the quality of life; the destabilization of the ecological economy; and the disruption of social order. In addition, each country has become vulnerable to the developments which take place in other parts of the global village, which are, of course, outside its own individual control.

Human society is increasingly recognizing these threats. Throughout the world huge resources, indeed hundreds of billions of U.S. dollars, are being spent annually to counteract them. While these efforts are to be praised because they prevent a part of the potential damage, it is nonetheless the case that on the whole these initiatives have reached a kind of stalemate – the destabilising factors now prevail and the scale of possible catastrophes is increasing rapidly.

The study-week focused on the question: how can we use our knowledge of chaos and self-organization to understand, predict and control such developments? To this end, the Academy brought together experts from the fields of mathematics and theoretical physics who study the general properties of chaotic systems, in addition to experts on a wide range of specific kinds of crises and disasters. The synergy of these fields of expertise has been highly successful in the past and the study-week explored further applications of such a joint-approach. Discussion took place on the moral, ethical and spiritual dimensions of proposed scientific initiatives and their implementation at the level of public policy. Altogether thirty-five world famous experts were brought together, representing the natural sciences, the social sciences, epistemology, and public policy. They came from the countries of the West and the East, and from the North and the South of the globe. Probably never before had such a variety of expertise been brought together for the purposes of a professional brainstorming discussion which was unconstrained by formal limitations. It focused on what can be done rather than simply drawing attention to the growing threats with which we are now faced.

In examining the problems of sustainability over the next decades, the study-week discussed such subjects as the shortfalls in global food supply; the deterioration of bio-diversity; the possible lack of response of the global village to the inevitable changes in the environment; the inefficient use of water resources; climatic changes; the economic burden of nuclear arma-

ment; and the threats and opportunities generated by the globalisation of the economy.

The study-week discussed a range of "instant" critical phenomena and addressed itself in particular to geological and geotechnical disasters and to the globally co-ordinated prediction of major earthquakes; socio-economic collapse in urban areas; political and economic crises, and their prediction; the outbreak of nuclear war; the use of electromagnetic terrorism and its threat to modern systems of communication and control; the self-organization of clusters of neurons and how this may help us to construct a rather general new approach to the control of critical phenomena; and the dangerous deficiencies in decision-making caused by greed and ignorance. In the search for a potential contribution by non-linear science to these problems, the meeting discussed scenarios of transition to a critical phenomenon. Such scenarios emerge when a chaotic process is examined at a not too detailed level (in the way that an oil painting cannot be understood through a microscope). They happen to be partly 'universal' and this is something which is shared by somewhat differing processes. The mathematical modelling of such scenarios opens up the possibility of predicting critical phenomena and may even supply the key to how to control them.

The work in this field requires high level professionals and the question of the education and training of experts is thus of crucial importance. For this reason the meeting discussed the very successful achievements in such training which has been obtained by the Abdus Salam International Centre for Theoretical Physics in Trieste.

On the positive side, the study-week may be seen as an attempt to reduce the Babel fragmentation which has taken place in the study of critical phenomena, more commonly known as disasters, catastrophes and crises. It outlined the common features of different critical phenomena and discussed commentary on each of them. It was also observed that in some cases it is possible to identify the future possible scale of a catastrophe as it approaches. In the case of many critical phenomena the work of the study-week amounted to a continuation of existing debate because a great deal of synergy in this field already exists. However, the meeting also identified the areas where the resources offered by non-linear dynamics have not as yet been exploited and stressed that certain groups of experts and scholars have not yet entered into dialogue. This may include the study of scenarios of transition to unsustainability and of specific ways of responding to it. Such a study may help to overcome a lack of interest in this question, something which was criticised by several speakers.

The step from many kinds of actual experience to a physical model is not easy, and the transition from a metaphor to an algorithm is always very

difficult. The synergy between observations on specific disasters and their general theory is neither a panacea nor an easy task. And the transition from a physical model which describes actual phenomena to knowledge which in a certain way is able to foresee or even predict them is even more difficult, and this is especially the case when one is dealing with human matters. Still, these new co-ordinated approaches provide hope that we may find as yet unexplored possibilities by which to overcome the present stalemate which exists in relation to how to face up to the many threats to our civilisation. This synergy probably reflects the paradigm formulated in *Fides et Ratio*: "human reason a capacity which seems almost to surpass its natural limitations. Not only is it not restricted to sensory knowledge, from the moment that it can reflect critically upon the data of the senses, but, by discoursing on the data provided by the senses, reason can reach the cause which lies at the origin of all perceptible reality" (§ 22).

In this way the human being becomes the vicar of God on earth within the advance of the creation in relation to the beings of nature from which he derives – and always with new methods – the means for survival and the achievement of growth and development, and this even in situations which are critical or apparently no longer sustainable. The immense range of the celestial bodies which pierce the firmaments – the sun, the moon, the stars, the galaxies, the comets – and the universal cosmic forces to be found on the earth, all have their laws which man must not change but which he must try to explore with his mind and employ for the purposes of his own survival and the attainment of growth and development. Therefore we should engage neither in the Casandra-like announcement of future catastrophes nor in an irresponsible optimism. Today, in the face of the global complexity of our contemporary context, the human being, more than ever before, is called upon to find that right kind of rationality (*orthòs logos*) which will achieve survival and sustainability through the application of new and well deployed practical criteria. For this reason, however limited the action of man within the cosmos may actually be, he is nonetheless a real participant in the power of God and must be able to build his own world, or rather an environment suited to his person integrated into his own space and his own special time.

V.I. KEILIS-BOROK
M. SÁNCHEZ SORONDO

THE PROGRAMME

This study-week will address the question: "What can basic science contribute to the survival and sustainable development of the world?" We will focus on the specific theme: possible applications of modern non-linear dynamics to the prediction and control of critical phenomena ("catastrophes") in nature and society. To this end, we will bring together experts in non-linear dynamics (mathematicians and theoretical physicists) and experts on the specific critical phenomena which are commonly recognized as serious threats to our world. We will also discuss the crucial problem of implementing scientific initiatives in public policy, and the moral, ethical, and spiritual dimensions of such initiatives.

Accordingly, the programme will consist of the following parts:

— *Problems of sustainability: response to threats of the time scale of decades.*
Responsible: Prof. P. Raven.

— *Problems of mankind's survival: response to the threat of catastrophes, which may happen at any moment.*
Responsible: Prof. G. Puppi, Prof. V. Keilis-Borok.

— *Scenarios of transition to critical phenomena.*
Responsible: Prof. N. Cabibbo, Prof. L. Pietronero.

— *Science and public policy.*
Responsible: Prof. G. Puppi, Prof. V. Keilis-Borok, Prof. P. Raven.

Premise

The problem. The world is facing major threats caused by the expansion of human activities, among them the deterioration of the environment, the depletion of natural resources, and the destabilization of economies and social order. The long-term threats (in the scale of decades) to the sustainability of our planet, like global warming, are accompanied by the immediate dangers of natural and man-made disasters; our vulnerability to them is greatly magnified with each passing year, and this undermines our ability to

maintain a sustainable and productive world into the twenty-first century and beyond.

Human society has increasingly recognized such threats. Throughout the world, huge resources, hundreds of billions of U.S. dollars, are being spent annually to counteract them. While these efforts are commendable because they prevent part of the potential damage, on the whole they have reached a kind of stalemate: the destabilizing factors prevail, and the scale of possible catastrophes is rapidly growing.

Both history and common sense tell us that basic research is pivotal to breaking such a stalemate. Indeed, since ancient times, basic science has repeatedly rescued humanity, providing "new solutions to old problems". The present study-week will engage in the search for such new possibilities, focusing on the major responsibility of today's scientific community.

At the same time, we recognize that scientific initiatives can be useful only if they can be implemented as a public policy and are acceptable to society from moral, ethical and spiritual points of view. These issues are also included in the programme.

How will this study-week differ from the escalating multitude of scientific meetings, from technical discussions to global forums, that are already dedicated to these problems? The distinctive features of this study-week will be:

— *brainstorming discussion* without any formal limitations;
— *a small number and a high level of participants*;
— *a focus on cutting-edge basic research*;
— *a focus on what can be done*, rather than simply alerting the audience to growing threats.

In combination these features are unique, so that this study-week will not duplicate other meetings devoted to similar subjects, but will complement them.

Why the Pontifical Academy of Sciences? Since hardly anybody else has the capacity to set up such an unusual meeting, with a potential for affecting the global agenda, we believe that the Academy has a responsibility to do so.

The Theme

As the specific theme for this study-week, we propose to focus on the applications of modern non-linear dynamics to the prediction and control

of critical phenomena. In order to develop that theme, we will bring together:

— experts in general studies of critical phenomena which employ methods developed mainly in mathematics and theoretical physics;

and

— experts on specific critical phenomena ("catastrophes" or "crises") that are encountered in nature and society.

The synergy that has been developed between these fields of expertise has been highly successful in addressing critical phenomena of various kinds. The present study-week will explore new untapped possibilities by which to apply this approach to the areas covered by the programme.

The study-week will also discuss the closely related issue of how to implement scientific initiatives at the individual level, and at the level of society as a whole. Implementation would require (i) an appropriate public policy and (ii) consideration of the moral, ethical, and spiritual problems involved in the implementation of initiatives based solely on scientific considerations.

V.I. KEILIS-BOROK

LIST OF PARTICIPANTS

Prof. WERNER ARBER (Pontifical Academician): University of Basel, Department of Microbiology, Biozentrum - Klingelbergstrasse 70, CH-4056 BASEL (Switzerland).

Dr. WALLACE S. BROECKER: Columbia University, Lamont-Doherty Earth Observatory - 61 Route 9W/P.O. Box 1000 - 14 Geochemistry, Palisades, N.Y. 10964-1000 (U.S.A.).

Prof. LUIS A. CAFFARELLI (Pontifical Academician): The University of Texas at Austin, Department of Mathematics RLM 8.100 - Austin, TX 78712-1092 (U.S.A.).

Prof. NICOLA CABIBBO (President of the Pontifical Academy of Sciences): Università degli Studi di Roma "La Sapienza", Istituto di Fisica - P.le A. Moro 5, I-00187 ROMA (Italy).

Prof. BERNARDO M. COLOMBO (Pontificial Academician): Università degli Studi di Padova, Dipartimento di Scienze Statistiche - Via S. Francesco 33, I-35121 PADOVA (Italy).

Prof. PARTHA S. DASGUPTA (Pontifical Academician): University of Cambridge, Faculty of Economics and Politics - Sidgwick Avenue, Austin Robinson Building, CAMBRIDGE CB3 9DD (United Kingdom).

Prof. JEFF DOZIER (NASA rapresentative): University of California, Santa Barbara, Donald Bren School of Environmental Science and Management - SANTA BARBARA, CA 93106 (U.S.A.).

His Excellency Msgr. AGOSTINO FERRARI-TONIOLO: Former Permament Observer of the Holy See to the Food and Agriculture Organization of the United Nations (F.A.O.) - Rome, and Pontificia Università Lateranense, Istituto "Itrusque Iuris", Diritto del Lavoro Comparato e Internazionale e dell'Organizzazione Internazionale - Via G. Palombini 12, I-00165 ROMA (Italy).

Prof. VLADIMIR E. FORTOV: Russian Academy of Sciences - 32a, Leninski prospect, room 315, MOSCOW 117993 (Russia).

Dr. URIEL FRISCH: Observatoire de la Cote d'Azur, Dep. G.D. Cassini - BP 4229, F-06304 NICE Cedex 4 (France).

Prof. MICHAEL GHIL: University of California, Los Angeles, Department of Atmospheric Sciences, Institute for Geophysics and Planetary Physics - 405 Hilgard Avenue, LOS ANGELES, CA 90095-1567 (U.S.A.).

Dr. PETER H. GLEICK: Pacific Institute for Studies in Development, Environment, and Security - 654 13th Street, Suite 104, OAKLAND, CA 94612 (U.S.A.).

Prof. JOSÉ LUIS GOTOR: Università degli Studi di Roma "Tor Vergata" - Via Orazio Raimondo 18, I-00173 ROMA (Italy).

Prof. RAYMOND HIDE (Pontifical Academician): Jesus College - Trul Street, OXFORD OX1 3DW (United Kingdom).

Dr. CALESTOUS JUMA: Harvard University, Kennedy School of Government, Center for International Development & Belfer Center of Science and International Affairs - 79 John F. Kennedy Street, CAMBRIDGE, MA 02138 (U.S.A.).

Prof. VLADIMIR I. KEILIS-BOROK (Pontifical Academician): International Institute of Earthquake, Prediction Theory and Mathematical Geophysics - Warshavskoye sh. 79, Kor 2, RU-113556 MOSCOW (Russia).

Prof. JOEL L. LEBOWITZ: Rutgers, the State University of New Jersey, Center for Mathematical Sciences Research - 110 Frelinghuysen Road, PISCATAWAY, N.J. 08854-8019 (U.S.A.).

His Excellency Msgr. JAMES T. McHUGH, Coadjutor: Diocese of Rockville Center - 50 North Park Avenue, ROCKVILLE CENTER, N.Y. 11570-4184 (U.S.A.).

Prof. BENOIT B. MANDELBROT: Yale University, Department of Mathematics - 10 Hill House - POB 208283, NEW HAVEN, CT 06520-8283 (U.S.A.).

Sir ROBERT M. MAY: Office of Science and Technology - Albany House 94-98, Petty France, LONDON SW1H 9ST (United Kingdom).

Dr. RAJUL PANDYA-LORCH: International Food Policy Research Institute (IFPRI), 2020 Vision for Food, Agriculture, and the Environment - 2033 K Street, N.W., WASHINGTON, D.C. 20006-1002 (U.S.A.).

Prof. WOLFGANG K.H. PANOFSKY: Stanford University, Stanford Linear Accelerator Center - P.O. Box 4349, STANFORD, CA 94309 (U.S.A.).

Prof. M. ELISABETH PATÉ-CORNELL: Stanford University, Industrial Engineering & Engineering Management (IEEM) - Terman Bldg., STANFORD, CA 94305 (U.S.A.).

Prof. CRODOWALDO PAVAN (Pontifical Academician): Rua Alvares Florence, 298 (Butantan), 05502-060 SAO PAULO, SP (Brazil).

Prof. LUCIANO PIETRONERO: Università degli Studi di Roma "La Sapienza", Dipartimento di Fisica - Piazzale Aldo Moro 5, I-00187 ROMA (Italy).

Msgr. TULLIO POLI: Segreteria di Stato, V-00120 VATICAN CITY.

Prof. ALBERTO QUADRIO CURZIO: Università Cattolica del Sacro Cuore, Facoltà di Scienze Politiche - Largo A. Gemelli 1, I-20123 MILANO (Italy).

Prof. MIKHAIL I. RABINOVICH: University of California, San Diego, Institute for Nonlinear Science - 9500 Gilmar Drive, LA JOLLA, CA 92093-0402 (U.S.A.).

Dr. PETER H. RAVEN (Pontifical Academician): Missouri Botanical Garden - P.O. Box 299, ST. LOUIS, MO 63166-0299 (U.S.A.).

Dr. ANDREW W. REYNOLDS: U.S. Embassy - Via Vittorio Veneto 119/A, I-00187 ROMA (Italy).

Prof. ANDREA RINALDO: Massachusetts Institute of Technology, Department of Civil and Environmental Engineering - Ralph M. Parson Laboratory, Room 48-209, CAMBRIDGE, MA 02139 (U.S.A.).
and
Università degli Studi di Padova, Dipartimento di Ingegneria Idraulica, Marittima e Geotecnica - Via Loredan 20, I-35131 PADOVA (Italy).

Prof. GIORGIO SALVINI: Accademia Nazionale dei Lincei - Via della Lungara 10, I-00165 ROMA (Italy).

Msgr. Prof. MARCELO SÁNCHEZ SORONDO: Chancellor of the Pontifical Academy of Sciences, V-00120 Vatican City, and Professor of the History of Philosophy at LUMSA University, Via della Traspontina 21, I-00193 ROMA (Italy).

Dr. YUKIO SATO: National Institute of Science and Technology Policy - 1-11-39, Nagata-cho, Chiyoda-ku, TOKYO 100-0014 (Japan).

Prof. GASTON G.S. SCHABER: Centre d'Etudes de Populations, de Pauvreté et de Politiques Socio-Economiques - B.P. 48, 44 rue Emile Mark, L-4501 DIFFERDANGE (Luxembourg).

Prof. MIGUEL A. VIRASORO: The Abdus Salam Centre for Theoretical Physics - Miramare - P.O. Box 586, I-34014 TRIESTE (Italy).

ADDRESS OF THE PRESIDENT OF THE PONTIFICAL ACADEMY OF SCIENCES TO THE HOLY FATHER *

Holy Father,

Let me first of all express our deep-felt gratitude for being received in your presence on the occasion of the study-week of the Pontifical Academy of Sciences. On the occasion of the Plenary Session, last October, you expressed the importance of a continuing dialogue between the world of sciences and the world of philosophy and theology. This dialogue and collaboration is essential in order to face some of the more pressing needs of humanity. With the rapid transformation and growth of human society new problems arise, since this growth must necessarily happen on a planet which has large but finite resources. The question arises of the sustainability of the human use of the planet's resources.

Scientists feel a particular responsibility when analysing these problems and searching for possible solutions, since many of these problems arise as a consequence of the marvellous progress of the medical and physical sciences during the last century.

Science alone is, however, not enough. It is essential that human society in its entirety rise to an understanding of the new responsibilities which derive from its very growth and material progress. It is therefore for all of us of great comfort that on many occasions you have clearly spoken of the need for renewed respect for natural resources and of the need to seek a renewed harmony between humanity and our planet.

The present meeting on "Science for Survival and Sustainable Development" concludes a series of three study-weeks organized by the Pontifical Academy in order to study the interaction between man and his environment.

In the first of these, held last November, we reviewed the interactions between life and the planet, an equilibrium which is under stress as a con-

---

* The following address was delivered by the President of the Academy, Prof. Nicola Cabibbo, at the Papal Audience on 12 March 1999.

sequence of human activities which give rise to entirely new effects such as the ozone hole in the arctic regions or the rapid increase in the concentration of green-house gases in the atmosphere.

In the second study-week, held in January, we examined the prospects for satisfying the food needs of the developing countries during the next century, a period when the population of these countries will reach unprecedentedly high levels. Thanks to the great strides forward of agricultural sciences, embodied in the "Green Revolution", the possibility of producing enough food is now less of a concern then it was a few decades ago. Of greater concern is, however, the question of the availability of food to the poorer populations, especially, but not only, in Africa, which already experiences widespread malnutrition.

These two meetings collected at the Academy some of the best world experts in applied sciences, from metereology to oceanography, from agriculture to the soil sciences, and from medicine to nutritional research.

In the present meeting we take a different approach, for whose conception we are grateful to Professor Keilis-Borok, that of bringing together experts in the study of sustainability, bio-diversity, and natural catastrophes on the one hand, and scientists who have made important contributions to the development of basic science, on the other.

We are convinced that basic science has much to contribute to the understanding and perhaps to the solution of the problems of sustainable development and the well-being of humankind.

The interaction between life and the planet is determined by non-linear effects, where small causes can lead to disproportionate consequences: appearently minor variations in the pattern of agricultural practices can lead to desertification, while relatively small quantities of fluorine-chlorine chemicals injected into the atmosphere has led to the ozone hole in the polar regions.

In recent decades basic research has made great strides in understanding phenomena which are similarly discontinous. One example is the progress in the understanding of transitions between different states of matter. A difference in temperature of a fraction of a degree is sufficient to transform barren ice into life-giving water. A small difference in velocity is sufficient to change an ordered flow of water into a turbulent one.

Our knowledge of these phenomena, which are easily studied in the laboratory, and prove amenable to mathematical modelling, is rapidly advancing. We hope that this progress will prove useful in understanding the more difficult problems of the interaction between life and the planet, as well as in suggesting possible avenues for their solution.

Holy Father: in expressing our deep gratitude for this audience we wish to assure you of our dedication in the pursuit of our scientific investigations, as well as our convinction that the dialogue between science and philosophy which you advocate is even more necessary given the new problems which seem to emerge from human progress.

NICOLA CABIBBO

ADDRESS OF THE HOLY FATHER POPE JOHN PAUL II TO THE PARTICIPANTS OF THE STUDY-WEEK *

Mr President,
Your Excellencies,
Ladies and Gentlemen,

1. I am pleased to welcome you on the occasion of the study-week organized by the Pontifical Academy of Sciences on the theme of the contribution of science to world development. I thank your President for his kind words and I extend warm greetings to you all, assuring you of my appreciation of the service which you give to the human community. You have chosen to reflect on the serious risks facing the planet as a whole and, at the same time, to consider possible steps for the safeguarding of creation on the eve of the Third Millennium.

2. In today's world, more and more people condemn the increasing harm caused by modern civilization to persons, living conditions, climate and agriculture. Certainly, there are elements linked to nature and its proper autonomy, against which it is difficult, if not impossible, to struggle. Nevertheless, it is possible to say that human behaviour is sometimes the cause of serious ecological imbalance, with particularly harmful and disastrous consequences in different countries and throughout the world. It suffices to mention armed conflict, the

---

* The following address was delivered by His Holiness John Paul II on 12 March 1999. It was published in *L'Osservatore Romano* on 13 March 1999.

unbridled race for economic growth, inordinate use of resources, pollution of the atmosphere and water.

3. Man has the responsibility of limiting the risks to creation by paying particular attention to the natural environment, by suitable intervention and protection systems considered especially from the viewpoint of the common good and not only of viability or private profit. The sustainable development of peoples calls on everyone to place themselves "at the service of all, to help them to grasp this serious problem in all its dimensions, and to convince them that solidarity in action ... is a matter of urgency (*Populorum Progressio*, 1). Unfortunately, economic and political considerations and arguments frequently override respect for the environment, making the life of peoples impossible or placing them at risk in some parts of the world. In order that the world may be habitable tomorrow and that everyone may find a place in it, I encourage public authorities and all men and women of good will to question themselves about their daily attitudes and decisions, which should not be dictated by an unlimited and unrestrained quest for material goods without regard for the surroundings in which we live, and which should be capable of responding to the basic needs of present and future generations. This attention constitutes an essential dimension of solidarity between generations.

4. The international community is called to cooperate with the different groups concerned, to ensure that the behaviour of people, very often inspired by exaggerated consumerism, does not disrupt economic networks, natural resources or the safeguarding of the balance of nature. "Mere accumulation of goods and services, even for the benefit of the majority, is not enough for the realization of human happiness" (*Sollicitudo Rei Socialis*, 28).

Similarly, the concentration of economic and political strength corresponding to special interests generates power centres which frequently act to the detriment of the interests of the international community. This situation leads to arbitrary decisions against which it is

often difficult to react, thus exposing entire groups of people to serious harm. Parity and balance require research and decisions to be carried out with transparency, with the aim of serving the common good and the human community.

More than ever, it is important that a political, economic and legal order be established, based on clear moral principles, so that international relations will have as their objective the promotion of the common good, avoiding the manifestations of corruption which seriously damage individuals and peoples, and not tolerating the creation of unfair privileges and advantages which favour the richer countries and social groups, economic activities developed without regard for human rights, financial paradises and regions exempt from the rule of law. Such an order should have enough authority with national bodies to intervene on behalf of the most disadvantaged regions and to promote social programmes aimed solely at helping these regions to advance on the path of development. On this condition, man will truly be a brother of every man and a cooperator with God in the management of the created order.

5. All those who have a responsibility in public life are also called to develop professional and technological training, and to implement training periods, especially for young people, enabling them to take an active part in national growth. Likewise, it is essential to train managers for developing countries and to carry out technological transfers towards these countries. This promotion of social balance, founded on the sense of justice and effected in a spirit of wisdom, will ensure respect for people's dignity, enable them to live in peace and enjoy the goods produced by their land. Furthermore, a well-organized society will be able to respond more rapidly to catastrophes which occur, in order to give assistance to peoples, especially the poorest and consequently most deprived.

6. Your efforts to work out reliable projections constitute a precious contribution to ensure that individuals, especially those who

have the responsibility of guiding the destiny of peoples, fully assume their responsibilities to future generations, removing the threats arising from negligence, gravely mistaken economic or political decisions, or lack of long-term planning.

The strategies to be adopted, as well as the necessary national and international measures, should have as their primary aim the well-being of individuals and peoples, so that all countries will enjoy "a wider share in the benefits of civilization" (*Populorum Progressio*, 1). By means of an equitable sharing of the funds allocated by the international community and low-interest loans, it is important to promote initiatives based on impartial solidarity, capable of supporting correctly targeted activities, a concrete application of the best adapted technologies and research corresponding to the needs of local peoples, thus ensuring that the fruits of technological and scientific progress do not exclusively benefit major companies and the more advanced countries. I therefore invite the scientific community to continue its research to better discern the causes of the imbalances linked to nature and to man, in order to anticipate them and to propose replacement solutions for situations which become intolerable.

These initiatives should be based on a conception of the world which places man at the centre and respects the variety of historical and environmental conditions, making sustainable development possible, capable of responding to the needs of the entire population of the world. This is especially a question of having a long-term perspective in the use of natural resources, ensuring that present resources are not exhausted by irrational and uncontrolled intervention.

7. People sometimes have the impression that their individual decisions are without influence at the level of a country, the planet or the cosmos. This could give rise to a certain indifference due to the irresponsible behaviour of some individuals. However, we must remember that the Creator placed man in creation, commanding him to administer it for the good of all, making use of his intelligence and reason. From this, we can be assured that the slightest good act of a

person has a mysterious impact on social transformation and shares in the growth of all. On the basis of the covenant with the Creator, to which man is called to turn continually, everyone is invited to a profound personal conversion in their relationship with others and with nature. This will enable a collective conversion to take place and lead to a life in harmony with creation. Prophetic actions, however slight, are an opportunity for a great number of people to ask themselves questions and to commit themselves to new paths. Consequently, it is necessary to ensure that everyone, particularly young people who desire a better social life in the midst of creation, is educated in human and moral values; it is also necessary to develop every person's social sense and attentiveness to others, so that all may realise what is at stake in their daily attitudes for the future of their country and the world.

8. At the end of our meeting, I ask the Lord to fill you with the spiritual strength needed to continue your efforts in a spirit of service to humanity and with a view to a better future on our planet. To all of you and to your loved ones I cordially impart my Apostolic Blessing.

SCIENTIFIC PAPERS

on

SCIENCE FOR SURVIVAL AND SUSTAINABLE DEVELOPMENT

I.

PROBLEMS OF SUSTAINABILITY: RESPONSE TO THE THREATS OF THE TIME SCALE OF DECADES

SUSTAINABILITY: PROSPECTS FOR A NEW MILLENNIUM

PETER H. RAVEN

Humanity stands at a defining moment in history. We are confronted with a perpetuation of disparities between and within nations, a worsening of poverty, hunger, ill health and illiteracy, and the continuing deterioration of the ecosystems on which we depend for our well-being However, integration of environment and development concerns and greater attention to them will lead to the fulfillment of basic needs, improved living standards for all, better protected and managed ecosystems and a safer, more prosperous future. No nation can achieve this on its own; but together we can – in a global partnership for sustainable development. (Agenda 21, Earth Summit; Sitarz, 1993).

The noted scientists and technologists who gathered for the Congress of Arts and Sciences in St. Louis in 1904 would not have understood the meaning of those ringing words. Instead of worrying about global inequities or the destruction of the environment, they were delighted with the prospects for a world in which the possibilities for progress seemed virtually unlimited. More than a century after the introduction of the steam engine, the fruits of the Industrial Revolution had become evident on every front, and the United States was looking forward to a future of international leadership. Taking my cue from the St. Louis Congress, I shall focus many of the following remarks on the role of the United States, but shall eventually broaden that view to encompass the world, a world in which the influence of the United States and other industrialized nations is pervasive.

When the delegates assembled in 1904, they would have been mindful of the death of Queen Victoria, who had given her name to an era that had witnessed the most extraordinary scientific, technical, and industrial advances that the world had known to that point. The St. Louis World's Fair

itself was celebrating not only the growing outreach and power of the United States, but also the broad vision of a diverse world that seemed to hold so much promise for the future. Theodore Roosevelt, the youngest American president, was in the White House, later to win the Nobel Peace Prize for his role in bringing about the end of the Russo-Japanese War; about to become the first American president to travel outside of the country, when he visited the construction site for the Panama Canal (completed in 1914); and poised, in defiance of Congress, to send the American Navy around the world as a show of national strength (1908).

At the same time as Americans were so excited about the prospects for the development of the airplane, of the automobile, of new modes of communication, and all of the other inventions that promised so much for the future, a few of them had also begun to realize that the world was not as unbounded and limitless as it once had seemed. Explorers had reached the far corners of the Earth, and knowledge was pouring in about its lands and its peoples: we increasingly knew what was there. Fredrick Jackson Turner, later to become America's preeminent historian, had announced the closing of the frontier, an idea that was to have great influence on collective visions of the world in the early years of the century. What were the turn-of-the-century antecedents of ecology, of sustainability, and of biodiversity – concepts that are now intellectual landmarks on the topography of the twenty-first century, but virtually unknown a hundred years ago?

The World Then and Now

At the turn of the century, after more than a hundred years of the Industrial Revolution, the global population stood at approximately 1.65 billion, with about 74 million people in the United States. In about four months, an event to be officially "celebrated" on October 12, 1999, there will 6 billion of us, including a billion added within the past 12 years, and the billion before that in 13 years. There are at present just over 270 million people in the United States. Human expectations have risen continuously over the course of the century, while the global population has more than tripled; consequently, the level of consumption in the industrialized world has risen to heights undreamed of just a few decades ago. Changes in the biosphere also have been unprecedented, with a major proportion of them having occurred during the past 50 years (Turner, 1990). Over this period, and for the past few hundred years, technologies have been invented and deployed, and the world has in what is geologically an instant of time been converted from a wild one to one in which human beings, one of an esti-

mated 10 million species of organisms, are consuming, wasting, or diverting an estimated 45 percent of the total net biological productivity on land and using more than half of the available fresh water, locally at rates that clearly cannot be sustained for long. The properties of the atmosphere have been and are being substantially changed by human activities, almost all major fisheries are under severe pressure, and habitats throughout the world have been decimated, with populations of alien plants and animals exploding and causing enormous damage throughout the world, while species extinctions have reached levels unprecedented for tens of millions of years. Despite the optimistic tone set by the Earth Summit declaration quoted above, with perhaps 3 billion additional people joining our numbers over the next half century, we will clearly have an increasingly difficult time in maintaining our current levels of affluence or in achieving the lofty goals which our historical progress seems to have made available to us. The scales and kinds of changes in the Earth's life support systems are so differenent from what they have ever been before that we cannot base our predictions of the future, much less chart our future courses of action, on the basis of what has happened in the past (Vitousek *et al.*, 1997).

As Bill McKibben has outlined in his book, "The End of Nature" (1989), we have arrived at a time when human beings are effectively managing the whole planet, for better or worse. The end of nature as he understands it is the end of nature functioning independently of human beings. This is the vision that was explicitly explored in the outstanding collection of essays, "Uncommon Ground" (Cronon, 1995). Specifically, in the field of conservation, those organisms that survive will do so because human beings manage the Earth's resources in such a way that this is possible; those that are lost will be lost for the same reason. The pressures we exert on global ecosystems are so extensive that their future is up to us. For these reasons, it has become clear that we clearly are living in the most difficult and challenging times that humanity has experienced. How did we get to this point, and what have been some of the warning signs along the way?

A mere 10,000 years ago, when crop agriculture was first developed at several widely scattered centers both in the Old World and the New, several million human beings, far fewer than the number of people who visit the museums of the Smithsonian Institution annually, populated the world, at about the density of Aboriginal peoples in Australia before European contact. The availability of larger quantities of food, on a more dependable basis that had existed before that time, created conditions for the rapid growth of the human population to an estimated 300 million at the time of Christ, a number that held more or less steady for a thousand years, grew to 1 billion around 1800, reached 2.5 billion by 1950 and will, as I men-

tioned above, reach 6 billion in the present year, 1999. As human numbers have grown, their impact on the environment have increased also, regional evidences of overgrazing or deforestation having been regarded with dismay by some people ever since Classical times. It has been during the period of the Industrial Revolution, from the mid-eighteenth century onward, however, that the evidence of widespread human domination of the natural environment has grown so rapidly and become so obvious as to affect the world view of every person concerned with the future.

The Growth of Environmental Consciousness: Before 1900

Following Columbus' landfall in the New World five centuries ago, at a time when the global population was less than a tenth of what it is now (about 500 million), waves of people from the Old World colonized the newfound lands and grew to great numbers and great power. The same phenomenon occurred throughout the world, as colonial expansion and the extension of often unsustainable forms of land use rapidly changed the face of the continents (Grove, 1995). As Andrews (1999, p. 18), put it, "Colonization ... was among other things an environmental policy". The ways in which relatively unspoiled lands were rapidly changed by the practices associated with colonization, and the ideal visions of such lands that persisted in the minds of Europeans far longer than they did on the ground, had a great deal to do with our collective understanding of the limited nature of local and ultimately global resources (McCormick, 1989). By the 1850s the problem of tropical deforestation was already being viewed as a problem on a global scale, and one that urgently demanded correction. Although less emphasized in the latter decades of the nineteenth century and the first half of the current one, the powerful metaphor of the destruction of Eden proved an enduring and influential one.

In Colonial America, the collective vision was one of an endless cornucopia of forests and meadows, rich in natural resources to be exploited – the destruction of the wilderness and the taming of nature were widely-accepted as desirable goals. The image of nature in all of its wonder and abundance, and the deep and abiding love of the land that Americans generally share, however, also have their roots in this early history: the land seems inexhaustible, rich, and nurturing beyond our wildest dreams. As Wallace Stegner (1980) put it, "While we were demonstrating ourselves the most efficient and ruthless environment-busters in history, and slashing and burning and cutting our way through a wilderness continent, the wilderness was working on us. It remains in us as surely as Indian names remain on the

land. If the abstract dream of human liberty and human dignity became, in America, something more than an abstract dream, mark it down at least partially to the fact that we were in subtle ways subdued by what we conquered". In these words, Stegner has captured the essence of the ethical, moral, and religious overtones to environmentalism, which are fundamentally important to our perceptions of the field, and underlie our hope of progress in the future. Although much of what we say and do is materialistic and operational, the reasons that we do it lie within ourselves.

Even in colonial times, some began to take seriously the evidence of threats to the bounty of the land, and to view the profligate use of natural resources as a problem (Nash, 1982; Shabecoff, 1993; Andrews, 1999). However, it was not until the advent of industrialization, roughly from the 1830s onward in America, that massive changes in the landscape began to become evident on many different fronts. In a relatively few decades, from the mid-nineteenth century onward, most of the prairies were cleared, the remaining great forests were cut, and farms and, increasingly, cities were established everywhere in the land – the activities noted so poetically by Stegner were carried on apace. In addition, it has gradually become clear that "nature" is a profoundly human construction: it can never be separated fully from our own values and assumptions (Cronon, 1995, p. 25).

Increasingly alarmed by these trends and their perceived effects on the future productivity of the land, influential writers and public figures, mostly living in the cities of the East, began to call for the preservation of some of our national wildlands, especially in the West: the sense of passing of the wilderness ultimately had a powerful effect on the national imagination. Ralph Waldo Emerson and Henry David Thoreau re-defined our relationship with nature, laying the foundation for modern environmentalism and the concept of sustainability. At the same time, Charles Darwin, by placing the human race clearly in the biological context of its evolutionary history, helped substantially to break down the dichotomy that had been so generally accepted earlier between people and nature. Subsequently, George Perkins Marsh, America's first true environmentalist, understood well the concept of the balance of nature and brought it to the attention of a wide public, basing his appreciation on his knowledge of his native state of Vermont, as well as on his wide travels in the Mediterranean basin and elsewhere; his 1864 book, *Man and Nature; or, Physical Geography as Modified by Human Action*, is a classic both of environmentalism and of ecology. Marsh saw clearly that the destruction of nature could not be sustained, and pointed out the need for care in the management of our resources for the sake of future generations. America's first national park, Yellowstone, was established the same year that Marsh's book was published. Another

notable and far-sighted early experiment in re-defining the relationship between man and nature was the establishment of the Adirondack Forest Preserve, later the Adirondack Park, by New York State, in 1885.

At the same time that concern about nature, and especially about the fate of the Western lands, was growing, another important trend was greatly influencing the development of environmentalism. The explosive growth of cities and the increasing urbanization of the population brought widespread urban pollution, along with the development of a new way of life that differed remarkably from that of the countryside: the same trend that had accompanied the advances of the Industrial Revolution earlier in England and elsewhere in Europe. The new urban-centered life, and the development of the many remarkable institutions that it made possible, provided an abundance that led to a growing equality and equity, but also gave rise to many new problems concerning the conditions under which people actually lived in those growing cities, swollen by the ranks of immigrants seeking a new life in America. Thus nearly 13 million immigrants came to the United States between 1890 and 1910, the great majority of them living in cities, where they were joined by large numbers of people moving from the farms. Coal dust, smoke, and toxic chemicals, open sewers, uncertain and often polluted water supplies, crowded and insanitary toilets – these were the commonplace experience of urban dwellers at the turn of the century. The collective realization of what the awful crowding in cities, the squalid living conditions and urban pollution meant to the lives of people became, along with the protection of natural resources, a second element of fundamental importance in the formation of American environmentalism (Andrews, 1999, chapter 7), one that ultimately contributed enormously to the strength of the modern environmental movement (Gottlieb, 1993).

The Science of Ecology

The essays that were presented by Oscar Drude and Benjamin Robinson in St. Louis in 1904 revealed an ecology that was in its earliest stages of development. Their papers were mainly concerned with plant distribution and the organization of plant communities around the world, with no reference to any of the dynamic concepts that have come to be associated with the modern synthetic science of ecology a century later. The term "ecology" had first been proposed by the German biologist Ernest Haeckel in 1866, but Haeckel had no particularly novel insights about the field. In developing the concept, he was referring to the web that linked organisms with their environment, an idea directly related to the notion of "natural history" as it

had been understood earlier. Essentially, the science of ecology is one that has developed entirely in the twentieth century.

At first, it was the study of plant ecology, and the relationships within plant communities that dominated ecology; but oceanography, limnology, and other disciplinary approaches now part of the field were developed during the same years. Efforts to chart the limits of plant distribution and to understand those limits in a historical sense were pursued actively, with terrestrial animal ecology coming along later (McIntosh, 1985). F.E. Clements, who had served as secretary for the ecology section of the 1904 St. Louis meeting, became an important pioneer and leader in the development of more dynamic concepts, and helped to lead ecology away from its roots as a purely descriptive discipline. During the same years, H.C. Cowles, at the University of Chicago, played a seminal role in the development of the science by adding his valuable insights to the concept of plant succession. Eventually, the British ecologist C. Elton in his book "Animal Ecology" (1927) laid the foundations for terrestrial animal ecology. There followed rapidly in the ensuing decades the development of quantitative community ecology as a field, and an appreciation of the dynamics of populations and of the relationships between populations in communities, the flow of energy and the movement of materials in communities (in the second half of the century), and finally the emergence of a science of systems ecology, in the development of which the American ecologists Eugene P. and Howard T. Odum played major roles.

Like all branches of science, ecology has become increasingly quantitative and theoretical, with an emphasis on mathematical modeling, population ecology, and feedback loops; scientists such as G. Evelyn Hutchinson and his student Robert MacArthur were important contributors in this area.

It needs to be emphasized at this point that ecology and environmentalism are by no means synonymous concepts: ecology is in fact a scientific discipline that deals with the relationships between organisms and with their environment and develops logical ways of examining and making predictions concerning them. A concept such as "sustainable development" is necessarily based on the principles of ecology, as those principles operate in a social and economic context. Notwithstanding this fundamental distinction, the development of the field of ecology into a strong scientific discipline during the course of the twentieth century is one of the factors of fundamental importance which allows us to evaluate the dilemma that faces us as we enter the new millennium. The whole set of biological relationships that it comprises provides the basis for understanding the reactions of different sets of populations, whether of humans or of other kinds of organisms, to their changing environment. Ecology likewise, especially through

the synthetic field of conservation biology, illuminates the fundamental principles on which our biological heritage can potentially be conserved for our future welfare.

Environmentalism in Twentieth Century America

Environmentalism in the United States was marked in the early years of the twentieth century by the emergence of the remarkable leadership of Gifford Pinchot, John Muir, and Theodore Roosevelt. These inspirational men considered in their individual ways that our natural resources should be managed so as to serve the needs of the future as well as those of the present: their influence was enormous, and persists to the present. They built particularly on the concept of parks and reserves, and that of safeguarding natural resources for all people. Among the events that marked the growth of environmentalism prior to World War II were the establishment of the National Audubon Society (1905), the controversy over Hetch Hetchy Valley in the Sierra Nevada of California (the valley was granted to San Francisco in 1913), the Migratory Bird Treaty Act established with Canada (1918), the establishment of the Civilian Conservation Corps (1933), and the passage of much federal legislation to regulate forests, water, and soil erosion during the 1930s. The influence of cartoonist Ding Darling (1876-1962; Lendt, 1979), who published widely syndicated and much-appreciated environmental cartoons from 1916 onwards, cannot be overestimated. As chief of the Biological Survey (later the Fish and Wildlife Service) in the 1930s, and because of his wide networking, he contributed a great deal to making Americans aware of their environment and what they were doing to it, and to the world – he clearly had an international vision of the environment, and projected that vision in many ways. And these are just a few samples of what was going on during those years.

During World War II, environmental concerns were largely sidetracked by the urgent ones associated with the war effort. Following the war, there occurred a period characterized by what Shabecoff (1993) called "careless optimism and materialism". Environmental concern gradually returned, however, as people were confronted on all sides with widespread evidence of severe problems. During these years, events such as the severe air pollution that occurred in Donora, Pennsylvania, in 1948, in which 20 people died and 14,000 became ill; the London "Killer Smog" that left 4,000 people dead in 1952; and concern over soil loss, water pollution, and the destruction of natural resources drew widespread attention and led to the enactment of new laws protecting people and natural lands.

One of the first books to call attention to these problems forcefully for a general audience was Fairfield Osborn's "Our Plundered Planet" (1948), which by its title as well as by its substance helped to stimulate serious and widespread debate. Osborn considered "the grand and ultimate illusion [to be] that man could provide a substitute for the elemental workings of nature". The concerns expressed by Osborn gradually moved to center stage in the public mind, his book having played a major role in stimulating concern about the environment and the directions in which we were heading.

Aldo Leopold, a great conservationist and philosopher, wrote some of the most stirring essays in the history of the field; his posthumously-published "A Sand County Almanac" (1949) has inspired generations of environmentalists. This book immediately became a landmark of the movement towards what we would now call sustainability, and is surely one of America's finest gifts to the world conservation movement, and thus to future generations. Leopold's "land ethic" speaks of a complex world dominated by human beings, who thus have either the power of good, nurturing care of their land, or the ability to degrade and destroy it. In his words, it "changes the role of *Homo sapiens* from conqueror of the land-community to plain member and citizen of it" (1943, p. 216).

Partly as a result of the writings of leaders such as Osborn and Leopold, and partly because of the increasing evidence of environmental degradation seen ever more widely, public concern over environmental matters reached new heights in the 1960s. The publication of "This is the American Earth", an exhibit-format book featuring the photographs of Ansel Adams and the poems of Nancy Newhall, by the Sierra Club in 1960, made a significant contribution to environmentalism and to a new way of thinking about the Earth at a spiritual level at the start of the decade. Over the following years, many influential writers and speakers began to warn of the dangers of excessive human domination of the Earth, generalizing from what had earlier been seen as individual, unconnected problems. They did so during a half century in which a world population that had grown by 850 million people during the preceding 50 years to a record level of 2.5 billion continued to increase at accelerated rates to its present level of 6 billion people. Such growth, coupled with industrial expansion from 1945 onward and increasing expectations on the part of consumers, greatly increased the strains on all ecological systems in ways that had become widely evident by the 1950s and 1960s.

In 1962, the first excepts of Rachael Carson's "Silent Spring" appeared in The New Yorker, and our common vision of our relationships with our planet were permanently altered. Clearly the most important environmental book written in America, "Silent Spring" focuses on chemical pesticides,

but with clear vision charts the destruction that technology can bring if carelessly applied. Carson presents a vision of a future world in which intelligent people can create a sustainable world. By doing so in such a convincing way, she moved environmentalism permanently to the center of the American agenda. Another landmark work was published near the end of the decade, when Paul Ehrlich's best-seller "The Population Bomb" (1968) dramatized and made available for a wide public for the first time the problems associated with rapid growth in human population, in effect adding a new dimension to the environmental debates.

The gathering momentum of the environmental movement culminated on Earth Day, April 22, 1970, when some 20 million Americans, one of every ten people in the nation, massed to demonstrate their concern over the state of the environment. Environmentalism had emerged as a mass social movement, resonating with civil rights and the other major social movements of the day. Many new environmental groups had been organized, and they were growing rapidly along with others that had been in existence earlier. Starting with the National Environmental Policy Act, signed into law on January 1, 1970, the concerns of those who were attempting to lay the foundations for a sustainable future were embodied in our laws, followed by the passage of the Clean Air Act. The Environmental Protection Agency was created at the end of the same year; the Clean Water Act in 1972. Of particular significance was the establishment of the Endangered Species Act in 1973: the world's most comprehensive legislation dealing with the conservation of biological diversity.

Earth Day in 1990 was even more significant in demonstrating the degree to which environmentalism had pervaded every aspect of American society, from corporations to consumer life styles, and had become a force that could not again be disregarded in the formation of public policy. What it called into focus, however, was that even though the environmentalism that was so strongly expressed in the 1960s had resulted in the establishment of outstanding environmental legislation, these accomplishments were not enough. Human nature combined with a failure to appreciate the global environmental situation, based partly on wishful thinking – the desire to continue on with "business as usual" – has resulted in bizarre and distorted conclusions like those of Easterbrook (1995), or the ones found daily in much of the economic press. Taken at face value, the assertions presented in such works would lead one to believe either that world economics functions in a vacuum, or that the natural productivity of the Earth and its maintenance and healthy functioning is of no interest in calculating human futures. Evidently, relatively few people in positions of authority are willing

to deal with the shock that comes when the global scale of these problems is recognized. Yet it is patently true that economic growth can be sustained over the long run only in the context of care for the environment.

Global Environmentalism

On a world scale, the formation of the United Nations in 1946 and the subsequent development of the organization gradually led to an increasing emphasis on problems associated with the environment. In 1968, the International Conference of Experts for Rational Use and Conservation of the Biosphere met in Paris under the auspices of UNESCO, and became the first major international meeting to examine human impacts on the environment. From this conference came the Man in the Biosphere (MAB) program, which specifically called for new ways of considering this relationship, and implementing improvements in it.

Four years later, in response to environmental problems in the Baltic region, the 1972 United Nations Conference on the Human Environment was convened in Stockholm. Here, the Canadian Maurice Strong began his brilliant international environmental career, and, when acting as head of the secretariat, brought about a strong examination of the relationship between the environment and development that has dominated international considerations of this area ever since. Building in part on the concepts expressed by the microbiologist and conservationist Rene Dubos, the conference examined the conditions under which human beings could exist in harmony with the rest of nature. Dubos' famous admonition, "think globally, act locally", has greatly influenced environmentalists, and his role in developing the concepts examined at Stockholm was of seminal importance. At the conference itself, a memorable role was played by Indian Prime Minister Indira Gandhi, who stated, "the inherent conflict is not between conservation and development but between the environment and the reckless exploitation of man and the earth in the name of efficiency".

Among the outcomes of the Stockholm conference was the formation of the Governing Council for Environmental Programs, a body that changed the following year (1973) into the United Nations Environment Program (UNEP), with its headquarters in Nairobi, Kenya. Its global orientation has served the world well during the 26 years of its existence, with many solid accomplishments to its credit. Nonetheless, its status as an agency supported by voluntary contributions has tended to marginalize some of its themes and the conclusions of its deliberations, and many believe that a

more central role for the environment within the U.N. General Assembly would be an appropriate response to the world environmental situation as we prepare to enter the new millennium. The scope of the world's problems does indeed seem to cry out for such a solution.

Another, and very different, event of key significance in the elaboration of the concept of sustainability was the publication by the Club of Rome of "The Limits to Growth" (Meadows *et al.*, 1972). The study this book reports uses comprehensive mathematical models to develop its conclusion that if present trends in world population continued, the limits to growth on the planet would be reached within a hundred years; the underlying conditions could be changed to establish a condition of ecological and economic stability that would last for the indefinite future; but if the world's peoples decided to change these conditions, the sooner they began, the more effective their actions would be. The remarkable feature of this book was its presentation of a comprehensive global model in which the various environmental, social, and economic factors that affect the human future can be considered in context for the first time. Although the study was widely reviled, particularly in economic circles, for the details of its projections, the majesty of its vision is as impressive today as when it first appeared, and the kind of reasoning it made possible remains fundamentally important. No enduring vision of the world's future can fail to take into account the effects of population growth, of affluence (consumption per person), or of the use of inappropriate technology, all of which need to be addressed in achieving global sustainability.

In practice, however, the appearance of the book set off a strong debate between the "cornucopians", who believed that environmental threats are grossly exaggerated, and that we should continue on with business as usual, and those who hold that catastrophes of various kinds are either upon us or just around the corner. What is certain in this debate is that early and intelligent actions will be required if some of the directions we are pursuing are to be changed; and change them we certainly must.

In the preceding remarks, I have deliberately not emphasized the growth of the global environmental movement, which parallels in different ways and with various regional and national characteristics that of the American environmental movement. McCormick (1989) and others have done a good job of charting the growth of what has become the largest social movement in history. One need only consider phrases such as "Chernobyl", "Times Beach", "Brent Spar", and "the Rainbow Warrior" to understand how the concepts of global environmentalism have pervaded our collective consciousness, and why. Certainly this movement, from the grass-

roots up through organizations, will have a major role to play in the organization of our responses to the problems that we so evidently confront as we enter the new millennium.

Sustainability

In the history of the environmental movement, "sustainability" is a recent concept that has proved powerful in describing the different factors that bear on our future. In 1987, the World Commission on Environment and Development published "Our Common Future", a report on the global environment in a human context. This report, which was adopted by the U.N. General Assembly calls for sustainable development as "development which meets the needs of the present without compromising the ability of future generations to meet their own needs". In other words, it combines the need to protect natural resources with the improvement of living standards: ecological systems and human systems working in harmony with one another. Pointing out that the problems of the environment in relation to human development are well known, the Commission called for urgent action to address these problems and to set the world on a sound course for the future. The Commission produced a brilliant and well-reasoned report, with strong recommendations in most fields affected by sustainable development. To some extent, its conclusions were built into the declarations from the Rio de Janeiro meeting five years later, but the objectives it laid out so clearly are still to be fully met. Achieving economic growth while taking into sufficient account environmental and social realities is our common goal, but it is very difficult to achieve. Despite the strong emphasis given to his area in the recommendations of the Earth Summit at Rio (Sitarz, 1993), relatively little progress has been made. Why has this been the case?

Twenty years after the Stockholm conference, it had become obvious that the state of the environment had deteriorated greatly from its 1972 condition. The authors of "Limits to Growth" (Meadows *et al.*, 1992, p. 2) wrote in their new analysis, "Beyond the Limits". "Human society has overshot its limits, for the same reasons that other overshoots occur. Changes are too fast. Signals are late, incomplete, distorted, ignored or denied. Momentum is great. Responses are slow ... if a correction is not made, a collapse of some sort is not only possible but certain, and it could occur within the lifetimes of many who are alive today".

In that same year, 1992, and once again organized under the tireless and effective leadership of Maurice Strong, the 1992 World Conference

on Environment and Development in Rio de Janeiro re-emphasized and expanded upon these themes, and led to the development of several important international treaties, including one dealing with climate change and a second with the protection, sustainable use, and fair and equitable sharing of biological diversity. The Earth Summit was a success to some degree, with the vision articulated twenty years earlier at Stockholm now widely accepted, and the depth of the problems confronting humanity generally understood. In addition, the enhanced role of non-governmental organizations (NGOs) in the meeting was an important advance that suggests one of the fundamental ways in which change may occur in the future. In addition, the organization of the Business Council for Sustainable Development was another important theme of the meeting, and one that has grown subsequently. The replenishment of the Global Environment Facility (GEF; formed in 1991), a financial mechanism to help developing countries deal with global warming, biodiversity loss, the pollution of international waters, and depletion of the ozone layer, was one important step, and several groups established or given new mandates at the time of the Rio meeting are addressing problems of great importance. What the Earth Summit did bring into sharp focus, however, was the huge difference between the concerns of the governments of industrialized countries, a fifth of the world's population with a per capita income of more than $20,000 and a life expectancy of 75 years, with those of the developing countries, four-fifths of the world's people, with a per capita income of about $1,200 and a life expectancy of 63 years. Some 1.3 billion people live in acute poverty, with incomes of less than $1 per day, 840 million of them receiving less than 80 percent of the U.N.-recommended minimum caloric intake, and thus literally starving.

When it became definite that India would attain independence, a British journalist interviewing Gandhi asked whether India would now follow the British pattern of development. Gandhi replied "It took Britain half the resources of the planet to achieve this prosperity. How many planets will a country like India require?". More recently, Wackernagel and Rees (1995) and others have emphasized again that if everyone lived at the standard of industrialized countries, it would take two additional planets comparable to Earth to support them, three more if the population should double; and that if worldwide standards of living should double over the next 40 years, twelve additional "Earths". Aspirations to such a standard of living are clearly unattainable, and yet advertising continually tells everyone that it is both appropriate and achievable. Even those who already live in rich countries continually strive to seek to improve their standards of living. The paradox pre-

sented by these relationships can be solved only by achieving a stable population, finding a sustainable level of consumption globally, accepting social justice as the norm for global development, and developing improved technologies and practices to make sustainable development possible.

We certainly understand better than ever the nature of the problems confronting us, but our willingness to deal with them, as we enter the new millennium, remains very limited, whether they be global warming, the destruction of forests, toxic pollution, the control of nuclear arms, or the destruction of the biological diversity on which we so confidently hope to base so much of our future prosperity. Seven years after the Earth Summit, industrialized nations have not funded the important recommendations of Agenda 21, the principal document that emerged from the meeting, and seem less interested in taking those recommendations seriously as time goes by. The lack of leadership by the United States, the world's wealthiest nation, has meant that the aspirations and plans developed in Rio de Janeiro in 1992 have mostly not been realized. How then can we and those who come after us expect to enjoy the benefits of a peaceful, healthy, and prosperous world in the twenty-first century and beyond?

Our collective inability, or perhaps unwillingness, to deal with conditions in the poorer parts of the world, on the one had, and the consumption patterns and lifestyles in more affluent parts of the world, on the other, pose serious obstacles to the attainment of global sustainability. With four-fifths of the world's people sharing the benefits of only 15 percent of the world's economy and their countries home to less than a tenth of its scientists and engineers, it is clear that the global system will operate properly only if there are increased financial contributions from the North. In most of the South, environments are deteriorating rapidly, and for large areas, the conditions in which people live are clearly unacceptable and unstable, often leading directly to environmental degradation (Shabecoff, 1996). Perhaps, as Shabecoff outlined, we are on the verge of a new enlightenment about the environment, but there are few indications that this is in fact the case.

Even though future societies based on information seem to promise less environmental degradation, the world view that so many of us share seems an unsuitable one for building a sustainable world. As Kai Lee (1993, p. 200) puts it, "how much misery will it take to make a global norm of sustainability first visible, then credible, then feasible, then inevitable? We do not know. And we do not know if the lessons of environmental disaster can be learned in time to ward off still more suffering. However bleak that prospect, we in the rich nations must bear the certain knowledge that our societies are both historically responsible for many of the circumstances

that imprison the poor and that we will on average fare much better than they. Against this background it is possible to see that sustainable development is not a goal, not a condition likely to be attained on earth, as we know it. Rather, it is more like freedom or justice, a direction in which we must strive, along which we search for a life good enough to warrant our comforts".

Biodiversity

The word "biodiversity" was coined by Walter G. Rosen at the U.S. National Research Council in 1986, in connection with the organization of a National Forum on BioDiversity sponsored by the U.S. National Academy of Sciences and the Smithsonian Institution (Wilson, 1988). Although it was a contraction of the familiar phrase "biological diversity", the new term took on an expanded meaning, and as Takacs (1996) points out, has become the rallying cry currently used by biologists and others to draw attention to the global ecological crisis broadly. At the 1986 conference, we were still largely dealing with a concept of "biological diversity" that tended to connote the array of species in the world, our knowledge of them, and the degree to which they were threatened by extinction. In contrast, "biodiversity", includes not only the genetic variation of those species but also all the ways in which they interact with one another in communities and ecosystems – the entire fabric of life on Earth. Viewed in this broader way, biodiversity becomes the stuff of sustainable development, our primary hope for sustainable management of the planet in the future, and, of course, the resource on which we hope to base the coming "age of biology" over the decades to come. In other words, a concept that started as "biological diversity", transformed into "biodiversity", has added to its original connotation of a set of individual organisms a much broader social meaning. In that sense, it approaches the meaning of earlier broad concepts such as "wildlife" or "nature".

It is notable that the formation of the Society of Conservation Biology occurred in the same year (1986) as the original conference on biodiversity. Like the conference itself, the formation of the Society signaled the maturity of an interdisciplinary effort in which the strands had been coming together for a number of years. An increasing maturity, based to some extent on the concepts that had been presented so poetically and well by Aldo Leopold 40 years earlier (Takacs, 1996), had deepened and broadened the conservation movement and the ways in which we can aspire to nurture the land and its living creatures.

The immediate inspiration for the formation of the concept of biodiversity was the sense of loss presented so clearly by authors such as Paul and Anne Ehrlich and Norman Myers in the 1970s and 1980s. Arguments based on the economic value of individual species, which are unquestionable and need not be elaborated here; those based on the value of ecosystem services, which in turn depend on interactions between species; and fundamental moral and ethical values, all play important roles in explaining the loss of biodiversity, estimated to amount to two-thirds of the species on Earth by the end of the coming century (Pimm and Brooks, 1999). Without biodiversity, we cannot respond well to the challenges we face, including global climate change: how will we form the new productive and stable biological systems of the future? A habitable planet requires the maintenance of the living systems that support all living things on Earth, including human beings.

Current extinction rates are several hundred times higher than those that have prevailed for tens of millions of years, and habitat destruction continues apace, so that extinction rates of 1,000 to 10,000 times those that existed in the past will wipe out species at a rate that has not prevailed since the end of the Cretaceous Period, some 65 million years ago – at just the time when humanity bases so much of its future hopes on its ability to use those species for human benefit. Furthermore, we have charted only a small fraction of the Earth's biodiversity, perhaps 1.6 million eukaryotic species even given a name, of an estimated total number of perhaps 10 million, with next to nothing on a global scale really known about such critically important groups as bacteria, fungi, and many groups of marine organisms. What we are losing, we do not even know: and perhaps never will.

A View of the Future

Over the course of the twentieth century, it has become overwhelmingly apparent that humanity cannot expect a healthy, peaceful, and productive future – in other words, a sustainable one – if we continue to live off the Earth's capital, rather than its interest: natural productivity. A world in which people are using or wasting nearly half of the total terrestrial photosynthetic productivity, one in which more than half of the available fresh water is already appropriated for human use, one in which the characteristics of the atmosphere are being altered rapidly, and one in which the species on which we hope to base the construction of sustainable and productive systems at the level of individual species and that of communities are disappearing in huge numbers – such a world will not be able to con-

tinue with its profligacy much longer without severe crashes of major ecological and economic systems (Meadows *et al.*, 1992). Global security likewise depends ultimately on environmental sustainability rather than on the expenditure of a huge proportion of the world's economic output to fund armies for rich, industrialized nations and poor ones alike (Myers, 1995). Food security, health, social justice – all are dependent on rising above our parochial and perhaps ingrained views of how to live, and learning together how to manage our planetary home for our common benefit. Empowering women throughout the world, seeking means to raise their status, and alleviating their poverty – microcredit has proved an effective strategy in this important effort – constitute among the most important actions to be taken to achieve sustainable development. Science and technology need to be fully applied in our striving toward global sustainability (Lee, 1993), but they alone will clearly not be enough. The new Social Contract for Science called for so forcefully by Lubchenco (1998), one in which scientists will address the most urgent needs of society; communicate their knowledge and understanding widely in order to inform society's decisions; and exercise good judgment, wisdom, and humility, constitutes a powerful call to action in a world that needs such action badly.

As the century comes to its end, it seems clear that the regulation of economic policy, with allowances for supporting the actions of the private sector, will have more impact on the environment than direct legislative initiatives. Conservative economists and radical environmentalists agree that the true value of the materials that we are using must become the basis of the sustainable commerce of the future, and that irrational taxes that drive unsustainable activities by mis-stating the value of their materials should be abandoned. Indeed, Myers and Kent (1998) have estimated that perverse subsidies leading to the destruction of natural resources worldwide amount to some $1.5 trillion annually, approximately twice as large as total global military spending, and larger than the economies of all nations on the Earth except the five largest – recognizing the undesirable nature of these subsidies and eliminating them or changing them in ways that will contribute to the sustainability of global ecosystems and resources would be one of the most important actions that humanity could take as we enter the new millennium. Perhaps the world's major corporations could in their own interest pursue an agenda in which the actual prices of resources were taken into account. In the design and construction community, for example, architects and building scientists are just now starting to operate by the rules of such an agenda, conserving energy and using new life cycle analysis (LCA) software tools to evaluate the environmental costs, such as resource depletion,

greenhous gas emissions, and energy consumption, of materials from "cradle to grave". Green consumerism is growing rapidly, with more than 31 million certified acres supplying "green" wood products in 1999. In addition, and of great importance, national and global systems of green accounting to reflect the full environmental costs of economic activities would help.

By pursuing strategies of the sort just reviewed, it might actually be possible to improve the potential condition of the world, and humanity's partly hard-wired tendency to behave as if we were still highly dispersed hunter-gatherers, rather than members of a rapidly growing human race comprising six billion people, some very rich, but many living in abject poverty. How could we build the political will to accomplish this? In view of the failure of the United States and other leading industrialized countries to address responsibly the agenda proposed at the Earth Summit in Rio de Janeiro in 1992, we cannot legitimately enter the new millennium with a sense of optimism. Despite this, we must be as effective as we can for the sake of those who will follow us, and we have significant choices to make that will clearly influence the shape of the world in the future, as analyzed effectively by Allen Hammond (1998).

Concretely, we could continue to strive to move sustainability closer to the center of the United Nations agenda, where it would be recognized as the most powerful factor in determining human futures. The United States could ratify the Convention on Biological Diversity, and all parties could refocus its activities on its three key objectives, which will help to conserve biodiversity and improve livelihoods, rather than allowing it to be consumed by questions of gene technology that have at best a marginal bearing on the survival of species around the world. The reform of the activities of the Convention, and their redirection towards appropriate objectives, would be a major step forward in the field of sustainable development. A global plan for the preservation of species, properly funded, would result in the greatest gift that we could possibly give to our descendants.

On the other hand, it may be that the model of a world driven by nations and the kinds of international institutions that were established in the wake of World War II will not prove to be dominant in the future. On the one hand, there is growing evidence that enlightened corporations are increasingly realizing that understanding and working with the conditions of sustainable development is a necessary prerequisite for success in the corporate world of the future (Hawken, 1993). John Browne, CEO of BP, for example, has set the company on a course that will embrace alternative energy sources and energy conservation, reasoning that in the face of global warming, they must do this if they are to continue to be a profitable energy

company in the future. How much more likely BP is to prosper than companies that ignore the conclusions about climate change that are so evident to the scientific community? Ray Anderson, chairman of Interface, an Atlanta-based carpet manufacturer, is likewise reorganizing his company's efforts around the conditions of the future, where sustainability will be a necessary condition of successful business, rather than those of the past. There are signs that the forestry and fisheries industries are starting to take sustainability seriously, and indications that consumers will increasingly demand appropriate certification for such products because of their concern for the environment. If corporations listen carefully to their stakeholders and take care to operate sustainably, they will affect the actions of governments and international agencies significantly and help to create conditions for their own prosperity, and for the world's sustainability. Frameworks such as that developed by The Natural Step, a Swedish organization that is having much influence throughout the industrialized world, will provide convenient blueprints to help guide us along the path of sustainability – but Kai Lee's (1993) principle that sustainability can perhaps best be viewed as an ideal, like justice, should be kept carefully in mind as we travel in that direction.

The kinds of grassroots activities that are promoting sustainability on a local basis have become a powerful force throughout the world: perhaps they are fundamentally only a re-emphasis of what has been traditional. Whether establishing local clinics and sustainable industries in the Biligiri Rangan Hills of South India, people-based ecotourism centers in native lands in Kenya, rebuilding a broken landscape at the Bookmark Biosphere Reserve in South Australia, learning how to ranch sustainably on the vast grasslands of the Malpai Borderlands of New Mexico and Arizona, or simply rooting out alien plants on Albany Hill in the San Francisco Bay Area, the people who are pursuing sustainability in a direct and personal way will hugely affect the shape of the world in the future. Outstanding books like those by Baskin (1997) and Daily (1997), explaining in detail how nature works and how we benefit from it in ways that most of us never consider will continue to play an important role in stimulating our desire to achieve sustainability. For example, watershed protection, the determination of local climates, and the protection of crops by birds and beneficial insects, including pollinators, that live in the ecosystems surrounding them are examples of ecosystem services – goods that nature provides without charge if we maintain sufficiently the integrity of the ecosystems that support them. In the light of this awareness, growing numbers of people will find ways to consume less energy, to recycle their materials, to participate in

the political process, to promote the acceptance of international understanding as a prerequisite for sustainability, and to support others, individually or in organized groups, who are pursuing these objectives.

For the basic conditions of change must clearly come from within us. A small minority of the Earth's residents cannot continue to consume such a large majority of the Earth's potentially sustainable productivity. By doing so, they will untimely destabilize their own future, as well as the futures of all other people. Population, overconsumption (among others, Schor, 1998, offers a powerful analysis of overconsumption in America), and the use of appropriate technology must all be brought into the equation if our common objective is to achieve a sustainable world in the new millennium. As Paul Hawken (1993) has put it so well, we need completely new ways of thinking about our place on Earth and the ways in which we relate to the functioning of natural systems if we are to find a better way to live in harmony with nature. Nothing less than a new industrial revolution (Hawken, Lovins, and Lovins, 1999) and a new agriculture (Conway, 1997) are required to make possible the sustainable world of the future. The task is incredibly challenging, but it is nonetheless one that we muct undertake if we responsibly understand the realities of our situation, and for the enduring good of those who come after us. It is also a fundamentally spiritual task. As Cronon (1955, p. 90) put it, "if wildness can stop being (just) out there and start being (also) in here, if it can start being as human as it is natural, then perhaps we can get on with the unending task of struggling to live rightly in the world – not just in the garden, not just in the wilderness, but in the home that encompasses them both".

In the words of Gandhi, which are most appropriate as we chart our course for the new millennium, "the world provides enough to satisfy everyman's need, but not everyman's greed". These words illustrated why Wilson (1993) was able to conclude that humanity would be able to overcome its drive to environmental domination and self-propagation with reason – why, in short, we are not necessarily suicidal in our approach to the world. In the spirit of Gandhi, one of the greatest leaders of our century, let us take his thoughts to heart and find the new inspiration that we so badly need at this incredibly challenging time. Global arguments may have little impact on the behavior of individuals unless they perceive the crisis as unbearably severe, something that impinges on people's lives in dramatic and frightening ways. By then it will be too late. Our ethics and our values must change, and they must change because we come to understand that by changing we will be happier people, guaranteeing a decent future for our children on a healthier planet in a more vibrant democracy in better neighborhoods and communities.

Many of the world's life-support systems are deteriorating rapidly and visibly, and it is clear that in the future our planet will be less diverse, less resilient, and less interesting than it is now; in the face of these trends, the most important truth is that actual dimensions of that world will depend on what we do with our many institutions, and with the spiritual dimensions of our own dedication. Clearly, the opportunities that are available to us now are very much greater than those contemplated with such joy by those who gathered in St. Louis in 1904, and the stakes are much higher.

Acknowledgments

I wish to thank for useful discussions in connection with the preparation of this paper, Robert Archibald, Gretchen C. Daily, Anne H. Ehrlich, Paul R. Ehrlich, Paul Farber, Kate Fish, Chris Hammer, Paul Hawken, Peter Jutro, Robert W. Kates, Jonathan Lash, Jane Lubchenco, Harold Mooney, Warren Muir, Norman Myers, Kerry ten Kate, George M. Woodwell.

REFERENCES

Adams, A. and N. Newhall (1960): *This is the American Earth* (Sierra Club Books, San Francisco).

Andrews, R.N.L. (1999): *Managing the Environment, Managing Ourselves. A History of American Environmental Policy* (Yale University Press, New Haven and London).

Baskin, Y. (1997): *The Work of Nature. How the Diversity of Life Sustains Us* (Island Press, Washington, D.C.).

Carson, R. (1962): *Silent Spring* (Houghton Mifflin Company, Boston).

Conway, G. (1997): *The Doubly Green Revolution. Food for All in the 21st Century* (Penguin Books, London).

Cronon, W. (ed.) (1995): *Uncommon Ground. Toward Reinveting Nature* (W.W. Norton & Company, New York).

Daily, G.C. (ed.) (1997): *Nature's Services. Societal Dependence on Natural Ecosystems* (Island Press, Washington, D.C.).

Easterbrook, G. (1995): *A Moment on the Earth: The Coming Age of Environmental Optimism* (Penguin Books USA, Inc., New York).

Ehrlich, P.R. (1968): *The Population Bomb* (Ballantine, New York).

Elton, C. (1927): *Animal Ecology* (Sidgwick and Jackson, London).

Gottlieb, R. (1993): *Forcing the Spring. The Transformation of the American Environmental Movement* (Island Press, Washington, D.C.).

Grove, R.H. (1995): *Green Imperialism. Colonial Expansion, Tropical Island Edens and the Origins of Environmentalism, 1600-1860* (Cambridge University Press, Cambridge).

Hammond, A. (1998): *Which World? Scenarios for the 21st Century. Global Destinies, Regional Choices* (Island Press, Washington, D.C.).

Hawken, P. (1993): *The Ecology of Commerce* (HarperCollins Publishers, New York).

Hawken, P., A. Lovins, and H. Lovins (1999): *Natural Capitalism. Creating the Next Industrial Revolution* (Little, Brown, New York).

Lee, K.N. (1993): *Compass and Gyroscope. Integrating Science and Politics for the Environment* (Island Press, Washington, D.C.).

Lendt, D.L. (1979): *Ding. The Life of Jay Norwood Darling* (Iowa State University Press, Ames).

Leopold, A. (1949): *A Sand County Almanac* (Oxford University Press, New York).

Lubchenco, J. (1998): 'Entering the Century of the Environment: A New Social Contract for Science', *Science*, 279, pp. 491-497.

McCormick, J. (1989): *Reclaiming Paradise* (Indiana University Press, Bloomington and Indianapolis).

McIntosh, R.P. (1985): *The Background of Ecology. Concept and Theory* (Cambridge University Press, Cambridge).

McKibben, W. (1989): *The End of Nature* (Random House, New York).

Meadows, D.H., D.L. Meadows, and J. Randers (1992): *Beyond the Limits* (Chelsea Green, Post Mills, VT).

Meadows, D.H., D.L. Meadows, J. Randers, and W.W. Behrens III (1972): *The Limits to Growth* (Universe Books, New York).

Myers, N. (1995): *Ultimate Security. The Environmental Basis of Political Stability* (W.W. Norton and Company, New York and London).

Myers, N. with J.V. Kent (1998): *Perverse Subsidies: Tax $s Undercutting Our Economies and Environments Alike* (International Institute for Sustainable Development, Winnipeg, Manitoba, Canada).

Nash, R. (1982): *Wilderness and the American Mind* (3rd ed. Yale University Press, New Haven and London).

Osborn, F. (1948): *Our Plundered Planet* (Little Brown, Boston).

Pimm, S.L. and T.M. Brooks (1999): 'The Sixth Extinction: how Large, how Soon, and Where?' in Raven, P.H. and T. Williams (eds.), *BioDiversity* (National Academy Press, Washington, D.C.).

Shabecoff, P. (1993): *A Fierce Green Fire. The American Environmental Movement* (Hill and Wang, New York).

P. Shabecoff (1996): *A New Name for Peace. International Environmentalism, Sustainable Development, and Democracy* (University Press of New England, Hanover and London).

Schor, J.B. (1998): *The Overspent American. Why We Want What We Don't Need* (HarperCollins Publishers, New York).

Sitarz, D. (ed.) (1993): *Agenda 21. The Earth Summit Strategy to Save Our Planet* (EarthPress, Boulder, CO).

Stegner, W. (1980): 'Coda: Wilderness Letter', in *The Sound of Mountain Water*, pp. 147-8 (University of Nebraska Press, Lincoln).

Takacs, D. (1996): *The Idea of Biodiversity. Philosophies of Paradise* (The Johns Hopkins University Press, Baltimore and London).

Turner, B.L., II. (ed.). (1990): *The Earth as Transformed by Human Action. Global and Regional Changes in the Biosphere over the Past 300 years* (Cambridge University Press, Cambridge).

Wackernagel, M. and W. Rees (1995): *Our Ecological Footprint: Reducing Human Impact on the Earth* (New Society Publishers, Gabriola Island, BC, Canada).

Wilson, E.O. (ed.) (1988): *BioDiversity* (National Academy Press, Washington, D.C.).

Wilson, E.O. (1993): 'Is Humanity Suicidal?' in *The New York Times Magazine*, May 30, 1993, pp. 24-28.

World Commission on Environment and Development (1987): *Our Common Future.*

FRESH WATER IN THE TWENTY-FIRST CENTURY: A SUSTAINABLE VISION

PETER H. GLEICK

INTRODUCTION

It seems inevitable that humans – barring some unforeseen catastrophe – will require a larger share of the Earth's limited renewable fresh water in the future than we do today. As the new millennium approaches, almost six billion people use nearly 30 percent of the world's total accessible renewable supply of water to grow food, run industries and cities, cool power plants, and meet other needs for drinking, cleaning, cooking, and playing. As the world's population climbs to seven, eight, or nine billion, more and more water will be required to satisfy basic needs and social, cultural, and economic desires. This water will come at an increasing financial and ecological price.

Perhaps the only certainty for anyone looking ahead is that the future is uncertain, unpredictable, and complex. But our present predicament – a consequence of our current policies, technologies, and institutions – is clear. The world community has failed to meet one of the most basic of human rights – the access to a basic amount of clean water. Today, billions of people lack access to sufficient water for their most basic of needs. As a direct consequence, millions of people die every year from a wide range of water-related diseases. Other severe water problems face us as well. Aquatic ecosystems and fisheries worldwide are being degraded and destroyed. Political disputes are flaring up over water that crosses political borders, and there is growing national and international competition for water in water-short regions. Despite increasing interest and concern over these problems, most of them have been getting worse over the past few decades, not better. As a result, while our water future can be only dimly seen, there is ample evidence that we will not like where we appear to be heading.

Decisions made every day about economic policies, technological choices,

and institutional structures all affect how we use resources to accomplish desired goals. By explicitly elaborating on what society needs and wants we can set goals for water use that are both sustainable and achievable. This paper argues that water is a basic human right guaranteed by international convention, law, and state practice. A basic water requirement for every human is proposed to meet this right. The paper also suggests that sustainable water use can be both defined and achieved, and that we already have the tools and policies that can move us in the right direction.

Background: Water in Crisis in the Twentieth Century

A wide range of ecological and human crises result from inadequate access to, and the inappropriate management of, freshwater resources. These include destruction of aquatic ecosystems and the extinction of species, millions of deaths from water-related illnesses, and a growing risk of regional and international conflicts over scarce, shared water supplies. As human populations continue to grow, these problems are likely to become more frequent and serious. New approaches to long-term water planning and management that incorporate principles of sustainability and equity are required. At the most basic level, a human right to water should be acknowledged by States. By declaring a human right to water and expressing the willingness to meet this right for those currently deprived of it, the world water community would once again place water on the top of the international sustainability agenda and would take a major step toward meeting one of the most vital requirements for human development and well-being. At a practical level, this paper explicitly explores concepts of the sustainable use of water. Seven "sustainability criteria" are discussed here, as part of an effort to reshape long-term water planning and management. Among these principles are guaranteed access to a basic amount of water necessary to maintain human health and to sustain ecosystems, basic protections for the renewability of water resources, and institutional recommendations for planning, management, and conflict resolution.

The twentieth-century water-development paradigm, which was driven by an ethic of growth powered by continued expansion of water-supply infrastructure, has been slowed in most industrialized nations as social values and political and economic conditions have changed. And while there have been efforts to extend this traditional development paradigm to many other parts of the world, massive water projects in developing countries are increasingly viewed with skepticism.

In the past, the primary goals of water policy were to support increasing levels of economic development and to figure out ways of increasing the

availability of fresh water to meet anticipated demands. Incidental to, or excluded from, these policies has been consideration of basic human needs, ecological water requirements, the roles of communities and culture, and the desires and needs of future generations.

The goal of relying on new supply projects to meet unlimited growth in demand has produced decidedly mixed results. Much of the concrete water infrastructure developed over the past 100 years has permitted great expansions of irrigated land and crop production necessary to feed rapidly growing populations. Massive urban population growth in most regions has been enabled by moving huge amounts of water from distant sources to cities. Devastating floods in many countries have been captured, curtailed, and tamed by flood control projects. The severe impacts of deep droughts are often mitigated by large storage systems that permit multi-year carryover of water.

Against these benefits must be weighed the full economic, social, and environmental costs of such projects and the apparent failure to provide for the basic water needs of billions of people. A focus on water supply has led to the neglect of attention to water use, leading in turn to many inefficient technologies and applications and to inequitable allocations of limited water supplies.

Perhaps most importantly, traditional approaches to water planning have neglected the ecological and environmental impacts of projects, both singular and cumulative. As a result, unanticipated or ignored ecological impacts have occurred, with sometimes devastating consequences. Among the kinds of ecological problems encountered are acidification of waters, unsustainable fisheries management, the wide spread of non-native species, and a cascade of biological effects from interbasin transfers and dam, reservoir, and aqueduct construction. Deforestation, urbanization, and agricultural chemical contamination also adversely affect aquatic systems.

Moving Toward Sustainable Water Policies: Water as a Basic Human Right

Access to a basic water requirement should be considered a fundamental human right supported by international law, declarations, and approved state practice. Providing that water should be the most important priority for governments, aid agencies, and communities (Gleick, 1999). By declaring a human right to water and expressing the willingness to meet this right for those currently deprived of it, the world water community would once again place water on the top of the international sustainability agenda and would take a major step toward meeting one of the most vital requirements for human development and well-being.

The term "right" in this paper is used in the sense of genuine rights under international law, where there is a duty on States to protect and promote those rights for an individual. The question of which rights are human rights has, of course, generated a substantial body of literature, as well as many organizations and conferences. Far less well explored, however, is the extent to which environmental rights are either found in, or supported by, existing human rights treaties, agreements, and declarations.

McCaffrey (1992) and Gleick (1999) tackled the legal background from the perspective of the UN (and related international law) human rights framework. Their analyses conclude that international law, international agreements, and the evidence of the practice of States strongly support the human right to a basic water requirement. They both conclude that there is a right to sufficient water to sustain life and that a State has the due diligence obligation to safeguard those rights as a priority. McCaffrey (1992) further argues that the devastating consequences of being denied such water should require that relevant provisions of existing humans rights instruments "ought to be interpreted broadly, so as to facilitate the implementation of the right to water as quickly and comprehensively as possible".

If we accept that there is a human right, to what extent does a State have an obligation to provide that water – and how much water – to its citizens? The international declarations and formal conference statements supporting a right to water do not directly require that States are responsible for fulfilling individuals' water requirements. Rather, States are obliged to provide the institutional, economic, and social environment necessary to help individuals to progressively meet those needs. In certain circumstances, however, individuals are unable to meet basic needs for reasons beyond their control, including disaster, discrimination, economic impoverishment, age, or disability. In such cases, the State must provide for basic needs (Gleick, 1996). The concept of a basic water requirement is discussed below.

Concepts of Water Sustainability

To broaden water policy to include issues of sustainability, a new debate has now begun, as reflected by the nature of the statements coming from the 1972 Stockholm Conference on the Environment, the 1977 Mar del Plata water conference, the 1992 Dublin statement, Chapter 18 of Agenda 21 from Rio, and more recent missives from the World Bank, the Global Water Partnership, and others (see Lundqvist and Gleick, 1996).

These statements suggest that incorporating characteristics of sustainability and equity in water planning and policy goals has become a major policy priority. Despite uncertainty about how best to define sustainability,

the basic principles require us to place a high value on maintaining the integrity of water resources and the flora, fauna, and human societies that have developed around them. And sustainability suggests that the costs and benefits of water-resource management and development be distributed in a fair and prudent manner. Together, these goals represent a commitment to nature and the diverse social groups of the present and future generations.

New Approaches for Sustainable Water Use

Water-resource planning in a democratic society requires more than simply deciding what project to build next or evaluating which scheme is the most cost-effective. Planning must provide information that helps the public to make judgments about which "*needs*" and "*wants*" can and should be satisfied. Water is now recognized as a common good and community resource, but it is also used as a private good or economic commodity; it is not only a necessity for life but also a recreational resource; it is imbued with cultural values and plays a part in the social life of our communities. The principles of sustainability and equity can help bridge the gap between such diverse and competing interests (Gleick, 1998a).

Regional and global water planning must now address such questions as: how much water is needed to satisfy the domestic use of a family in a dense urban center or in a rural agricultural community? Should people be able to use as much water as they can pay for? Under what situations should water be delivered to farmers at rates below full operating and capital costs? How much water is needed to maintain ecological systems and environmental quality and services, and at what level? How much water should be available and at what quality for the use of future generations?

A set of criteria for guiding water-resource management is presented below. These sustainability criteria constitute an ethic that helps prioritize competing claims over water. The real challenge of this ethic is to define the specifics. What do sustainability and equity mean when applied in the real world? What kind of planning practices are consistent with these objectives?

While not all will agree with this specific approach, the direction that is set out can be used to guide rational and meaningful debate over water-resource policy. Rather than allowing the overall goals to be determined by the outcomes of fights among the most powerful and wealthy interest groups, goals to further a genuine common interest can be forged and real conflicts can be resolved in a fair and equitable manner based on democratic ideals. In the absence of democratic dialogue, water-resource development can only continue down a course plotted decades ago, one that may have been appropriate then, but which fails to meet the challenges of the next century.

With respect to water resources, as with many other resources, sustain-

ability has not been clearly defined, though several recent efforts have made progress in defining the issues (Koudstaal *et al.*, 1992; Plate, 1993; Golubev *et al.*, 1988; Gleick *et al.*, 1995; Raskin *et al.*, 1995; Gleick, 1998a). Water is not only essential to sustain life, it also plays an integral role in ecosystem support, economic development, community well-being, and cultural values. How are all these values, which sometimes conflict, to be prioritized? What is to be sustained? For how long? What are the benefits? Who are the beneficiaries? In the context of freshwater resources, any discussion of sustainable development requires that we understand the stocks and flows of global, regional, and local water resources, and the benefits or services that those resources must provide (Gleick *et al.*, 1995).

The simplest definition of the sustainable use of water would require the maintenance of a desired flow of benefits to a particular group or place, undiminished over time. Benefits involve cultural values and issues, and are a function of the stock of, and the demand for, water, both of which vary with technology and population. Demands for water include not just what people "need", but what they "want". This latter demand is potentially much larger than minimum basic needs (Gleick, 1996), described below. This simple definition of sustainability, however, would permit maintaining benefits to one user group at the expense of another user group. A better definition would incorporate the requirement that benefits to all current users be maintained, without reducing benefits to other users, including natural ecosystems. This definition is flawed too, by excluding explicit rights for future generations or growing populations.

Further refinement requires that the sustainability of current benefits be maintained without affecting the ability to provide comparable benefits into the future. The desired set of benefits provided by a resource does not have to be, and is unlikely to be, the same across different users or periods of time. Indeed, desired benefits of water use vary widely given political, religious, cultural, and technological differences. But in any realistic discussion of sustainability, the benefits to be provided must be explicitly evaluated. Benefits of water use can be sub-divided in several ways: by form or sector of use, such as domestic, agricultural, industrial, and ecosystem use; or by the well-being provided by use, such as economic wealth, human and ecological health, level of satisfaction, and so on. Sophisticated measures of well-being are often difficult to quantify but provide a more complete view of the consequences of resource use than the traditional measures of simple quantities of per-capita use.

Unsustainability of Water Resources

Gaining an understanding of the sustainable use of water can also be approached by understanding what constitutes an "unsustainable" use of

water. Using the definitions above, water use is unsustainable if the services provided by water resources and ecosystems, and desired by society, diminish over time. Equity also requires that a reduction of services over time to one user group be declared "unsustainable" even if other users are able to maintain their desired services. It should be noted, however, than inequities by themselves are not unsustainable – ironically, many inequities in resource allocation and use can be maintained for indefinite periods of time.

Unsustainable water use can develop in two ways: (1) through alterations in the stocks and flows of water that change its availability in space or time; and (2) through alterations in the demand for the benefits provided by a resource, because of changing standards of living, technology, population levels, or societal mores.

Water availability is affected by both natural and anthropogenic factors, including climatic variability and change, population growth that reduces per-capita water availability, contamination that reduces "usable" water supplies, physical overuse of a stock, such as groundwater overdraft, and technological factors. Similarly, demands for water are not constant; they increase with growing populations, change as social values and preferences change, and increase or decrease with technological innovation and change.

Two problems deserve special attention: increasing populations and changing technology: the first leads to both decreasing per-capita water availability and increasing overall demand; the second affects both water supply and demand. Assuming constant levels of total water availability, increasing populations lead directly to decreasing per-capita water availability and pressures on the levels of benefits or the mix of benefits that water provides. Ultimately, unlimited population growth must lead to decreasing water availability, the reallocations of water from one user or sector to another, the unsustainable "mining" of non-renewable stocks of water, and, in the end, decreasing overall benefits.

Technological developments can alter water availability, and can affect the amount of water required to satisfy demands. In theory, practically unlimited quantities of fresh water are available by mining water currently trapped in glaciers and icecaps, or on an even larger scale, through the mass desalination of seawater. In practice, however, increases in overall water supply should occur only where the value of water exceeds the economic and environmental costs of supplying that water.

Similarly, changes in technology can increase or decrease the amount of water required to supply a particular societal benefit. If technological development proceeds independently of water constraints, a new technology to supply energy, for example, may require more water than previous alternatives. If water resources are constrained, technology can be manipulated to

reduce overall water requirements in the same way that energy efficiency technologies reduce energy needs without sacrificing the desired benefit.

Finally, truly sustainable water use must involve the management of the distribution of water in space and time. Social systems – i.e., institutions – to control water resources must be capable of coping with changes in supply and demand and in responding to varying priorities of water use under different conditions.

A Framework for Sustainable Water Management and Use

Gleick *et al.* (1995) offer a working definition of sustainable water:

the use of water that supports the ability of human society to endure and flourish into the indefinite future without undermining the integrity of the hydrological cycle or the ecological systems that depend on it.

This definition provides an overarching framework by which decisions about human water use can be judged. To make decisions about how to allocate and use water resources, however, more detailed goals and criteria need to be identified. Explicit criteria and goals that lay out human and environmental priorities for water use are presented in Table 1.

Table 1. *Sustainability Criteria for Water Planning.*

A basic water requirement will be guaranteed to all humans to maintain human health.
A basic water requirement will be guaranteed to restore and maintain the health of ecosystems.
Water quality will be maintained to meet certain minimum standards.
These standards will vary depending on location and how the water is to be used.
Human actions will not impair the long-term renewability of freshwater stocks and flows.
Data on water resources availability, use, and quality will be collected and made accessible to all parties.
Institutional mechanisms will be set up to prevent and resolve conflicts over water.
Water planning and decision-making will be democratic.

The criteria and goals of Table 1 are the result of considerable dialogue and analysis with academic, governmental, and non-governmental interests working on regional, national, and international water problems. They are not, by themselves, recommendations for actions; rather they are endpoints for policy – they lay out specific societal goals that could, or should, be attained. In particular, these criteria provide the basis for alternative "visions" for future water management and can offer some guidance for legislative and non-governmental actions in the future (Gleick *et al.*, 1995). In contrast, without specific criteria to guide planning, unsustainable water policies are inevitable.

Policy discussions must inevitably turn to identifying how much water is required to satisfy these priorities and which of the many economic, technical, educational, and regulatory means that are available should be pursued. While debate on how to attain these goals is unavoidable and desirable, having a set of clear targets will help focus the ultimate policy decisions.

Criterion 1. Meeting Basic Human Water Requirements

The first criterion listed above sets as a primary goal the provision of a basic amount of water for meeting the essential needs of humans. This elementary goal, common to many different interpretations of sustainability over the past few years, was raised in "basic needs" requirements of the 1977 Mar del Plata statement, restated in the United Nations Agenda 21, which explicitly recognized the standing of both humans and ecosystems, and is part of the compact for human development described in the 1994 United Nations Development Programme (UNDP) *Human Development Report*. For humans, insufficient access to potable water is the direct cause of millions of unnecessary deaths every year. The provision of a certain amount of fresh water to support the human metabolism and to maintain human health should be a guaranteed commitment on the part of governments and water providers.

A true minimum can only be defined for maintaining human or ecological survival. For humans, this amount is approximately five liters per person per day under average climatic conditions and levels of activity. Additional basic needs have been quantified, however, for providing sanitation services, preparing food, and bathing. Gleick (1996) recommends that a "basic water requirement (BWR)" standard be developed to satisfy these needs, and argues that an appropriate standard based on physiological, cultural, social, and technical factors is 50 liters per person per day of clean water (Table 2). No legal or institutional mechanisms exist, however, to guarantee even this basic requirement to present and future generations.

The first sustainability criterion, therefore, guarantees access to this basic water requirement to meet the fundamental domestic needs of people.

Table 2. *Basic Water Requirements for Human Needs* [a].

Purpose	*Recommended Basic* (liters per person per day)
Drinking Water [b]	5
Sanitation Services	20
Bathing	15
Food Preparation	10

[a] Excluding water required to grow food.
[b] This is a true minimum to sustain life in.

Criterion 2. Meeting Basic Environmental Water Requirements

The second of the criterion listed above requires a minimum amount of water be guaranteed to meet the essential needs of natural ecosystems. This goal was also supported as part of the "basic needs" requirements of Agenda 21 of the United Nations (UN, 1992). Some limited efforts have been made to set minimum requirements for certain threatened or high-priority ecosystems, but few criteria have been set, particularly in the developing world.

In part because of the lack of clearly defined legal water rights, many aquatic ecosystems and individual species have become severely threatened or endangered. The recent disasters to befall the natural fisheries of Lake Victoria and the Aral Sea are but two examples. Overall, more than 700 species of fish have been recognized by international organizations as threatened or endangered. In just the last couple of years, many more have been added to the list because of increasing pressures on water resources, including several anadromous species. Anadromous fisheries, in particular, are extremely vulnerable to changes in water supply and quality and to modifications in habitat (Covich, 1993; Nash, 1993; NRC, 1996).

While efforts are being made to identify basic ecosystem water requirements, there is little agreement about minimum water needs for the environment and few legal guarantees for environmental water have been set. The ecosystems for which water is necessary include both natural ecosys-

tems where there is a minimum of human interference and ecosystems that are already highly managed by humans. Societal decisions will have to be made regarding the degree to which these ecosystems should be maintained or restored and the indicators by which to measure their health. Examples of such decisions include identifying stretches of undisturbed rivers to preserve, establishing minimum flow requirements, reallocating water from major water projects to the environment, and developing standards to protect wetlands and riparian habitats. Protecting natural aquatic ecosystems is not only vital for maintaining environmental health, but there are important feedbacks between these systems and both water quality and availability as well. Two examples of the new focus on ecosystem water needs are the recent decision to place a cap on further development and diversions in the Murray-Darling river system in Australia (MDBMC, 1996) and the complete revision of South African water law to include water for ecosystems as a fundamental priority (MWAF, 1996; Gleick, 1998b).

Ultimately, allocations of water for the basic needs of ecosystems will have to be made on a flexible basis, accounting for climatic variability, seasonal fluctuations, human needs, and other factors. Management will have to follow an adaptive model where decisions are to be reviewed frequently based on the latest information and special efforts are made to avoid irreversible environmental consequences. More intelligent water resources management will be necessary to sustain our aquatic biological resources. In particular, we must maintain adequate water quantities and qualities for natural habitats, minimize alterations of natural ecosystem processes and losses of biodiversity and integrity, and preserve remaining natural freshwater habitats with high biodiversity and many endemic species.

Criterion 3. Water Quality Standards

Different uses require water of differing qualities. As a result, water-quality standards for different purposes must be developed and water quality must be monitored and maintained to meet these standards. Water in most developed countries is protected from contamination by national regulations (World Health Organization, 1984; United States Environmental Protection Agency, 1992; Minister of National Health and Welfare, 1992). These water-quality standards are supposed to ensure that potable water is reasonably free from contaminants known to affect human health. In many parts of the developing world, however, even minimal water quality standards are not in place, leading to widespread cases of waterborne diseases. Lack of sufficient, clean drinking water and sanitation services leads to many hundreds of millions of cases of water-related diseases and between

five to ten million deaths annually, primarily of small children (Table 3) (Nash, 1993; WHO, 1995; Warner, 1995).

Water used for non-human consumption need not be protected to the drinking water standards. For example, water used for many industrial, commercial, or landscaping purposes could be protected to a lower standard, with substantial economic savings. Similar water quality criteria need to be developed for ecological water requirements. Substantial effort should go into identifying these differences and developing ways of meeting various demands with water at appropriate levels of quality.

Table 3. *Estimates of Global Morbidity and Mortality of Water-Related Diseases.*

Disease	*Morbidity (episodes/year)*	*Mortality (deaths/year)*
Diarrhoeal Diseases	1,000,000,000	3,300,000
Intestinal Helminths	1,500,000,000 (people)	100
Schistosomiasis	200,000,000 (people)	200
Dracunculiasis	100,000 (people infected)	–
Trachoma	150,000,000 (active cases)	–
Malaria	400,000,000	1,500,000
Dengue Fever	1,750,000	20
Poliomyelitis	114	–
Trypanosomiasis	275	130
Bancroftian Filariasis	72,800,000 (people)	–
Onchocerciasis	17,700,000 (people)	40,000 (mortality caused)

Source: WHO, 1995.

Criterion 4. Renewability of Water Resources

Freshwater resources typically are considered renewable: they can be used in a manner that does not affect the long-term availability of the same resource. Renewable freshwater resources can be made non-renewable by mismanagement of watersheds, overpumping of groundwater, land subsidence, and aquifer contamination. Water policy should explicitly protect against these irreversible activities.

Groundwater stocks are renewable on timelines that depend upon the rate of inflow of water, the rate of withdrawals of water, and the geophysi-

cal characteristics of the aquifer. In some instances, overpumping of groundwater – the extraction of groundwater at a rate that exceeds the rate of natural recharge – can continue for some time with no adverse consequences if the aquifer is recharged during wet periods. Thus, a short-term non-renewable use may still be compatible with long-term renewability.

In regions where groundwater recharge rates are extremely low, such as in many arid and semi-arid regions, overpumping of groundwater is unsustainable and represents a one-time use of a resource stock – the same as pumping oil out of the ground. Eventually, the costs of taking out additional cubic meters of water will exceed their economic value to the user. This kind of water use is going on in several regions of the world (Table 4), including Saudi Arabia, Yemen, India, parts of the western U.S., and northeastern China, to mention only a few of the major problem areas.

Table 4. *Heavily Exploited Aquifers of the World.*

Region	*Aquifer*	*Average Annual Recharge (km^3/yr)*	*Average Annual Use (km^3/yr)*
Alegeria/Tunisia	Saharan basin	0.58	0.74
Saudi Arabia	Saq	0.3	1.43
China	Hebei Plain	35	19
Canary Islands	Tenerife	0.22	0.22
Gaza Strip	Coastal	0.31	0.50
United States	Ogallala	6 to 8	22.2
United States	selected Arizona	0.37	3.78

Source: Margat, 1996.

Some forms of groundwater pumping may lead to the irreversible decline in the ability of a region to store water in the ground. Even where overpumping during dry periods may, in theory, be replenished by rainfall during wet periods, geophysical characteristics may prevent this in practice. Excessive groundwater pumping in parts of the Central Valley of California, for example, has led to land subsidence, which reduces the ability of wet years to fully recharge groundwater aquifers. Estimates are that California's Central Valley has lost over 24 billion cubic meters of storage capacity owing

to compaction of over-exploited groundwater aquifers (Bertoldi, 1992). To put this loss in perspective, the entire storage capacity of all constructed reservoirs in the state is under 60 billion cubic meters (DWR, 1998). Over-pumping of ground water in coastal aquifers can also lead to irreversible and unsustainable effects, including salt water intrusion and the ultimate contamination of the entire groundwater stock.

Surface waters can also be contaminated or lost through watershed mismanagement. For example, animal grazing or excessive human use at high elevations can lead to fecal contamination of surface runoff in mountain streams. Urbanization can lead to storm runoff that is lost to sewers rather than feeding streams or recharging groundwater. Water managers and land-use planners must coordinate whenever these kinds of land-use decisions can lead to irreversible changes in the hydrological cycle.

Criterion 5. Data Collection and Availability

If water planning and management are to be democratic and effective, data on all aspects of the water cycle must be collected and made available in an unrestricted manner. At present, data on many aspects of regional and national water supply and use are not collected and when they are, are not widely available. At the extreme, some national governments continue to classify basic water data for so-called security reasons. This is unjustified and greatly inhibits effective water planning and management.

Substantial data gaps exist on the condition of different groundwater basins, extraction amounts, current pumping practices, and recharge rates. Similarly, water-use information is sketchy or site-specific, making actions for increasing efficiency or improving conservation programs hard to plan and implement. Information should be produced in reasonable time with reasonable resources, and it should be freely and widely shared.

Recent advances in electronic communications makes sharing resource information easy and inexpensive. In particular, Internet resources related to water are growing at a phenomenal rate, and many sources of information are already freely available (see, for example, www.worldwater.org). This trend should be encouraged and expanded.

Criteria 6 and 7. Institutions, Management, and Conflict Resolution

Criteria for sustainability must include more than biological or physical characteristics. They must also provide guidance for the institutions that are to resolve conflicts over water and deal with the unavoidable uncertainties and risks in decision making. The greatest debates over water in the past

several decades have focused on how to reach particular *goals*. The water debate must now be broadened to address the *means* by which these goals are set. Accordingly, sustainability criteria must also apply to water-resources management, particularly to ensure democratic representation of all affected parties in decision making, open and equitable access to information on the resources, and the options for allocating those resources.

Water planning and decision-making in many regions is limited to a narrow range of professionals trained in engineering, agriculture, and the hydrological sciences. The power of these groups remains significantly greater than that of rural interests, religious or ethnic minority groups, environmental groups, academics, and other users. Mechanisms to broaden their participation are needed. Ways must be found to incorporate and protect the interests of future generations – a fundamental criterion of sustainability as defined by the United Nations in Agenda 21 (UN, 1992).

In addition to mechanisms to broaden participation, institutional mechanisms need to be set up to prevent and resolve conflicts over water. There is a long history of conflict over shared water resources, described in detail in Gleick (1998b). Nearly half of the land area of the earth is part of an international river basin and more than 260 nations share water with a neighboring country. Although a wide range of tools for resolving water disputes already exist, their effectiveness varies greatly depending on the issue and the extent of political manipulation and interference. The most effective approach is specific treaties among river basin nations allocating water, setting up management oversight, and developing acceptable standards for operations and water quality. Unfortunately, few of the world's international rivers have such treaties, and many of the existing ones inadequately address either current or future problems (Wolf, 1997).

Another approach – the development of general international principles – has also been tried, with limited success. The International Law Commission has worked for many years to define such principles, and while much progress has been made, the application of these principles to solving specific regional conflicts has had very limited success (McCaffrey, 1993). Future institutions and efforts to settle the problems posed by international rivers must be open and democratic, and must resolve conflicts over water in an equitable, prudent, and fair manner.

Perhaps the greatest flaw with many water institutions is their failure to adequately address issues of equity. Equity is a measure of the fairness of both the distribution of positive and negative outcomes as well as the process used to arrive at particular social decisions. The sustainability goals in Table 1 explicitly incorporate institutional criteria for participation and conflict resolution so as to ensure at least a degree of procedural equity necessary for sustainability.

Some would argue that sustainability should be defined narrowly so that questions of equity are excluded. But from this perspective, sustainability could be achieved under otherwise morally reprehensible conditions. For example, the terrible health conditions in many parts of the world tied to inadequate water supplies (Table 3), are certainly "sustainable", but no ethical argument can be made for sustaining them. Similarly, higher rates of species extinction may be tolerated for some time, but the moral implications of failing to slow them must be addressed. Questions of equity overlap with sustainability when trying to determine what is to be sustained, for whom it is to be sustained, and who decides. In general, great disparities in wealth, inequities in power between women and men, and discrimination based on race, ethnicity, or age can lead to conflicts that undermine attempts to achieve sustainability. Thus, a fair political process is itself a necessary component of sustainability.

Summary and Conclusions

A communications and information revolution is sweeping the globe. There is renewed interest in reaching out to outer space. International financial markets and industries are increasingly integrated and connected. And efforts are being made to ensure regional and global security. In the context of these exciting human developments, our inability to provide the most basic water needs of billions of people may be remembered as the twentieth century's greatest failure.

It is time to declare that access to the most basic water requirements is a human right, protected by international law, declarations of governments and international organizations, and normal state practices. The right to water sufficient to meet basic needs should be an obligation of governments, water management institutions, or local communities. While in some regions, governmental intervention may be necessary to provide for basic water needs, many areas will be able to use traditional water providers, municipal systems, or private purveyors within the context of market approaches. Unfortunately, there are many reasons why governments or water providers may be unable to provide this amount of water, including rapid population growth or migration, the economic cost of water-supply infrastructure in regions where capital is scarce, inadequate human resources and training, and even simple political incompetence. Nevertheless, failure to provide this basic need is a major human tragedy. Preventing that tragedy should be a major priority for local, national, and international groups.

The sustainability criteria presented provide a framework for prioritizing competing interests and for making decisions about filture water use and management. The first two criteria require that we identify and meet basic allocations for humans and ecosystems, which are to be satisfied before other demands. In this respect, the approach described above defines criteria for "basic needs" as recommended by Agenda 21 of the United Nations. This paper presents the concept of a basic water requirement (BWR) for human domestic needs and recommends that a BWR for drinking, basic sanitation services, human hygiene, and food preparation be guaranteed to all humans as a fundamental human right. The work of international organizations, United Nations agencies, and individual researchers suggest that this right can be explicitly quantified at around 50 liters per person per day.

Hundreds of millions of people, especially in developing countries, currently lack access to this BWR, and this causes enormous human suffering and tragedy. Furthermore, rapid population growth and inadequate efforts to improve access to water ensure that this problem will grow worse in some areas before it grows better. This problem should be a far higher priority for governments, water providers, and international aid organizations than it appears to be.

The sustainability criteria not only set out quantity and quality requirements, but they also provide some institutional guidance. It is easier to agree and quantify minimum standards for human health, which has some biophysical basis, than it is to determine how much water should be allocated for irrigation or for industrial use, but these decisions need to be made as well. In allocating water to these other demands, planners must move beyond simple economics and incorporate concepts such as efficiency, equity, and participatory democracy as well.

The sustainability criteria are not meant to be all encompassing. They help answer only certain questions for public policy and planning. Nevertheless they can provide a strong set of guidelines for positive action. Ultimately, until discussions about the sustainable use of water become an integral part of long-term water planning, the world will be faced with continued unsustainable water use and threats to both human and ecological survival.

REFERENCES

Bertoldi, G.L. (1992): 'Subsidence and Consolidation in Alluvial Aquifer Aystems', in *The Proceedings of the 18th Biennial Conference on Groundwater, U.S. Geological Survey*, pp. 62-74.

Covich, A. (1993): 'Water and Ecosystems', in P.H. Gleich (ed.), *Water in Crisis: A Guide to the World's Fresh Water Resources* (Oxford University Press, New York), pp. 40-55.

Department of Water Resources (DWR) (1994): *California Water Plan Update, Final Bulletin*, pp. 160-93, Sacramento, California.

Gleick, P., Loh, P., Gomez, S., and Morrison, J. (1995): *California Water 2020: A Sustainable Vision*, Pacific Institute Report, Pacific Institute for Studies in Development, Environment, and Security, Oakland, California.

Gleick, P.H. (1996): 'Minimum Water Requirements for Human Activities: Meeting Basic Needs', *Water International*, 21, pp. 83-92.

Gleick, P.H. (1998a): 'Water in Crisis: Paths to Sustainable Water Use', *Ecological Applications*, Vol. 8, No. 3, pp. 571-579.

Gleick, P.H. (1998b): *The World's Water 1998-1999: The Biennial Report on Freshwater Resources* (Island Press, Washington D.C.).

Gleick, P.H. (1999): 'The Right to Water: A Basic Human Right', Paper in preparation.

Golubev, G.N., L.J. David, and A.K. Biswas (1998): Sustainable Water Development: Special issue. *Water Resources Development*, Volume 4, No. 2 (June).

Koudstaal, R., R.R. Rijsberman, and H. Savenije (1992): 'Water and Sustainable Development', *Natural Resources Forum*, pp. 277-290 (November).

Lundqvist, J. and P. Gleick (1996): *Sustaining our Waters into the 21st Century*, Report to the Comprehensive Global Freshwater Assessment of the United Nations. Stockholm Environment Institute, Stockholm, Sweden.

Margat, J. (1996): Comprehensive assessment of the freshwater resources of the world: Groundwater component, Contribution to Chapter 2 of the Comprehensive Global Freshwater Assessment, United Nations.

McCaffrey, S.C. (1992): 'A Human Right to Water: Domestic and International Implications', *Georgetown International Environmental Law Review*, Volume V, Issue 1, pp. 1-24.

McCaffrey, S.C. (1993): 'Water, Politics, and International Law', in P.H. Gleick (ed.), *Water in Crisis: A Guide to the World's Fresh Water Resources* (Oxford University Press, New York), pp. 92-104.

Minister of National Health and Welfare (MNHW) (1992): *Guidelines for Canadian Drinking Water Quality* (5th edition, Canadian Government Publishing Center, Ottawa, Canada).

Ministry of Water Affairs and Forestry (MWAF) of South Africa (1996): 'Fundamental Principles and Objectives for a New Water Law in South Africa', Report to the Minister of Water Affairs and Forestry of the Water Law Review Panel (January).

Murray-Darling Basin Ministerial Council (MDBMC) of Australia (1996): 'Setting the Cap: Report of the Independent Audit Group', Murray-Darling Basin Ministerial Council (November).

Nash, L. (1993): 'Water Quality and Health', in P.H. Gleick (ed.), *Water in Crisis: A Guide to the World's Fresh Water resources* (Oxford University Press, New York), pp. 25-39.

Plate, E.J. (1993): 'Sustainable Development of Water Resources', *Water International*, 18, pp. 84-94.

Raskin, P., E. Hansen, and R. Margolis (1995): *Water and Sustainability: A Global Outlook*, Polestar Series Report No. 4, Stockholm Environment Institute, Boston, Massachusetts.

United States Environmental Protection Agency (1992): Drinking Water Standards and Health Advisory Table, U.S. Environmental Protection Agency Drinking Water and Ground Water Protection Branch, San Francisco, California.

United Nations (1992): *Earth Summit Agenda 21: The United Nations Programme of Action from Rio*, New York, Chapter 18 of Agenda 21 is devoted to freshwater resources.

Warner, D.B. (1995): 'Water Needs and Demands: Trends and Opportunities from a Domestic Water Supply, Sanitation and Health Perspective', Presented at the Workshop on Scenarios and Water Futures, Stockholm Environment Institute, Boston, Massachusetts, 28-30 September 1995.

Wolf, A. (1997): 'International Water Conflict Resolution: Lessons from Comparative Analysis', *International Journal of Water Resources Development*, Vol. 13, No. 3.

World Commission on Environment and Development (WCED) (1987): *Our Common Future* (Oxford University Press, New York).

World Health Organization (WHO) (1971): *International Standards for Drinking Water*, Third Edition, World Health Organization, Geneva, Switzerland.

World Health Organization (WHO) (1984): *Guidelines for Drinking-Water Quality*, Volume 1, World Health Organization, Geneva, Switzerland.

World Health Organization (WHO) (1995): *Community Water Supply and Sanitation: Needs, Challenges and Health Objectives*, 48th World Health Assembly, A48/INF.DOC./2, 28 April, 1995, Geneva, Switzerland.

NITROGEN AND THE FUTURE OF WORLD AGRICULTURE

CRODOWALDO PAVAN and JOHANNA DÖBEREINER

"The human species is biologically an extraordinary success, precisely because its culture can change ever so much faster than its gene pool. This is the reason cultural evolution has become adaptively the most potent extension of biological evolution. For at least 10,000 and perhaps 1,000,000 years man has been adapting his environment to his genes more often than his genes to his environment". (Dobzhansky, 1962).

With these qualities, *Homo sapiens* has turned out to be the most influential species in the development and harmony of the biosphere, the environment which all species (*Homo sapiens* included) depend on for their survival.

In reality, instead of "the development and harmony" mentioned above, and in full respect and admiration for the achievement produced by our species, it would be more appropriate, in this case, to mention the "development and exploitation" of the biosphere, as revealed by the following statement of Raven (1998): "Human populations, having grown from fewer than 1 billion people at the start of the Industrial Revolution, or 2.5 billion in 1950, to nearly 6 billion today, are estimated to be consuming or wasting directly or indirectly approximately 45 percent of the total net terrestrial productivity – the total energy that enters biological systems on land, as a result of the activities of an estimated 300,000 species of photosynthetic organisms".

In other words, a single species, among more than 20 million others, uses for its convenience 45% of the total net terrestrial productivity, causing many problems as stated in the invitation we each received to participate in this study-week: "The world is facing major threats resulting from expansion of human activities, among them deterioration of the environment, depletion of natural resources, and destabilisation of economies and social order".

The important point here is that these facts are well known by a great number of enlightened people all over the world. This has resulted in many discussions, published papers, recommendations and resolutions being submitted by the population to governmental and sopragovernmental agencies. Unfortunately, however, very little is being done to correct or even to moderate the negative effect being imposed on the Earth's biosphere by the "superior" species.

Let us turn to the main subject of our talk, the role of nitrogen compounds in the biosphere. We shall mainly discuss food production, which presently is one important way by which certain nitrogen compounds enter into the biosphere in excessive amounts causing ecological troubles which should be highlighted, discussed and, of course, corrected whenever possible.

Food is one of the essential components for the maintenance of any living organism. Primitive humans obtained food by the process of hunter-gatherers following the general and natural rules of all wild animals, with some mental advantages over the other species. Evolving culturally, the human species passed through a herdsman or nomad stage before reaching what we call civilisation.

For about ten thousand years after the farming stage, it had a rather continuous and slow development and progress, with evident changes in the biosphere, which did not amount to anything that would cause much concern. With the advances of the so-called Industrial Revolution starting about 250 years ago, food production among other similar human activities started to be very rapid and more aggressive towards the environment, causing deep concern about the problems created for future generations.

In the years fifties and sixties, with the support of the Rockfeller Foundation, the Ford Foundation and several agencies and governments, a project was developed to improve cereal crop production. This resulted in what was called the Green Revolution, which from the outset was successful as a method to increase food production, as was well demonstrated in Mexico and India, between 1950 and 1970. However, the methods involved the abundant use of fertilisers, pesticides, herbicides and suitable machinery, causing at least two important problems. One was ecological; the excessive use of fertilisers and pesticides, which is a normal practice in agriculture today, which produces pollutant effects in the ecosystems with serious consequences. The other effect of social consequence was properly described by Raven and Johnson (1992): "The Green Revolution has resulted in increased food production in many parts of the world through the use of improved crop strains. These strains often depend on increased inputs of fertilisers, water, pesticides and herbicides, as well as the greater use of machinery, means that are not normally available to the poor".

The negative effect caused by the excess of nitrogen and phosphorus fertilisers introduced into the biosphere over a great part of the globe over the last decades, and discussed by many authors, prompted the Ecological Society of America to organise two panels of scientists to analyse the problem and present it to the public. This resulted in four publications: Vitousek P.M. *et al.* (1997, 1997a), *Human Alteration of the Global Nitrogen Cycle: Causes and Consequences*; and Carpenter S. *et al.* (1998), *Nonpoint Pollution of Surface Waters with Phosphorus and Nitrogen*.[1] The main concern of those documents is the excess of phosphorus and nitrogen delivered into the biosphere by human activities, mainly into the aquatic ecosystem, causing serious long-term environmental consequences for large regions of the planet.

The statements and the conclusions of these documents are of great importance and interest and are being circulated among interested scientists, governments and the public in general. Such groups should obtain the support of governmental agencies, philanthropic organisations and the general public, so that serious efforts can be made to correct this devastating situation which compromises the future of our species.

Although other agents and human activities play important roles in the increase of the amount of certain phosphorus and nitrogen compounds in the biosphere, as well stated in these documents, it seems that food production practices are principally responsible for this at the present time. One of the causes of the pollutant effect of fertiliser is that in the normal procedure for the use of industrial fertiliser in a crop plantation, the granulated solid fertilisers are distributed at random in the soil and the recommendation is that the quantity to be used should be the double of the one that the plant can absorb. The explained reason is that the ramified roots of the plants do not occupy the entire volume of soil available so that the recommended excess is to permit the plant to take sufficient fertiliser. Another cause is that during the time the fertiliser is in the soil, it is partly washed by irrigation or rain and transported to the aquatic system of the biosphere. The excess not absorbed by the plants continues to be washed after harvest and what is not washed away may be used either by wild plants or the next season's crop. Another habit that helps to increase the extra amount of these pollutant fertilisers in the soil is the belief of many farmers, especially the economically more prosperous ones, that the more fertiliser you use, the better will be the harvest. For this reason the amount used is often far in excess of the needs of the plants and the pollution is correspondingly

[1] These two panels organised by ESA are in reality new versions of a similar panel "Biological Nitrogen Fixation" organised by the US National Research Council, chaired by Prof. F. Hardy and published in 1996 (Hardy, 1996) as will be discussed later.

increased. A small part of the nitrogen fertiliser not used by the plants is converted into atmospheric nitrogen through a process of denitrification by the soil's bacteria.

Let us examine some important information provided by the report of the panels mentioned above (Carpenter *et al.*, 1998): "Between 1950 and 1995 about 600 million metric tons of phosphorus fertiliser was applied to the Earth's surface, primarily on croplands". After an analysis of what happened to the fertiliser, the plant and the crop, the finding is that, in that period, about 350 million metric tons, or roughly half of this phosphorus, has accumulated in the world croplands representing an average of 25% increase in the phosphorus content of agricultural soil, which corresponds to an average accumulation of 22 kilograms of surplus phosphorus per hectare each year. Part of this phosphorus in the croplands of the USA is washed and taken to large areas of lakes and reservoirs, destroying the biological equilibrium.

A similar or even worse situation occurs in relation to the release of nitrogen into the biosphere as seen in the results of the Vitousek *et al.* (1997) panel: "In fact, humans have already doubled the rate of nitrogen entering the land-based nitrogen cycle and that rate is continuing to climb". Or another statement: "During the past century, human activities clearly have accelerated the rate of nitrogen fixation on land, effectively doubling the annual transfer of nitrogen from the vast but unavailable atmospheric pool to the biologically available forms. The major sources of this enhanced supply include industrial processes that produce nitrogen fertilisers, the combustion of fossil fuels and the cultivation of soybeans, peas and crops that host symbiotic nitrogen-fixing bacteria".

Another phrase of interest is: "The amount of industrially fixed nitrogen applied to crops during the decade from 1980 to 1990 more than equated all industrial fertiliser applied previously in human history".

Both reports make it clear that if these errors are not corrected the benefits derived from the prevailing situation will far from compensate, practically and morally, the burdens being placed on the future generations.

To heat up our discussion, letus mention another alarming phrase used by Vitousek *et al.* (1997): "One study predicts that by the year 2020 global production of nitrogen fertiliser will increase from the current level of about 80 Tg[2] to 134 Tg per year". It is of interest to remember that in 1950 the global production was 3 Tg.

---

[2] Tg = Teragram, the standard unit of measurement for analysing the global nitrogen cycle and corresponds to one million metric tons or simply one million tons of nitrogen.

In both panels, the role of the excess of fertiliser and the products thrown into the biosphere are discussed, the negative results are well explained and the conclusion is that things must be changed if we want to secure an acceptable future for our species.

The process of food production by the present main sources is far from being able to provide sustainable development. The problem involves one of the most important activities of the human species, the food production essential for the species, whose present level will have to be increased in the future for two basic reasons: the increase in the population size and the no less important need for more food to satisfy the demands which will be made by the expected increase in the social level of at least 80% of world population.

There are many things to be done to correct these anomalous situations; we will deal in this talk with the possible substitution of the total, or at least partial, quantity of industrial nitrogen fertilisers by the bacterial biological nitrogen fixation (BNF), which is not a pollutant in nature.

This was also part of the programs organised by governmental American agencies in 1979. In that year, after the United Nations Conference on Science Technology for Development (UNCSTD), the US government, through its Agency for International Development (AID), initiated a series of programs on co-operation and training with developing countries. In at least three of these programs, projects on BNF were included.

In 1992 the US National Research Council appointed a panel to evaluate the contribution of the USAID supported BNF research conducted under these programs. We reproduce here a few of the suggestions and conclusions that are of important interest for the purpose of the present discussion:

a) "The panel concluded that expanded use of biological nitrogen fixation is equally critical to future crop and tree production in both developed and developing countries" and "accordingly, a major part of the report provided justification for expanding investments in BNF around the world".

b) "Growing concerns about the environment, energy, nutrition and agricultural sustainability make the need for BNF research even more compelling".

c) "Clearly, it is not realistic to consider sustainable agriculture on a broad scale in the absence of BNF; research is needed to optimize the contribution of BNF to sustainable agriculture".

The chapter entitled "Recommendation" begins with the following:

d) "Investment in Biological Nitrogen Fixation Research. Biological nitrogen fixation continues to be a high-priority research area, with expanding focus both for the developed and developing countries. There is a need

to relieve the more than $20 billion (and growing) cost of fertiliser nitrogen and its substantial environmental damage, coupled with the expanding requirement for food and the desire for more sustainable agriculture and forestry. These dictate research investment in BNF in the 1990s that is even more urgent than in the 1970s and 1980s. Research to expand our knowledge of nitrogen fixation and develop economic applications and management system for developed and developing countries must be pursued". Hardy 1996 and 1993.

Further arguments favouring a better use of BNF for agriculture are based on the discovery of the so-called endophytic nitrogen fixing bacteria in non legume plants. This new phase in microbiology started with the rediscovery of the genus Azospirilum by Döbereiner and Day (1975), a bacteria that they found living in the interior of gramineous plants. This was a completely new system, related to the well known nitrogen fixing bacteria that live in legume plants. For reviews see Boddey 1987 and 1995 and James and Olivares 1997.

A few years later, the introduction of a new method facilitated the isolation of nitrogen fixing bacteria through the use of the semi-solid malate medium, therely initiating an explosive phase of discoveries in the association of plants and nitrogen fixing bacteria, not only in Gramineae but also in a great number of other Monocotyledonous and Dicotiledonous families of plants. This was recorded historically in Döbereiner *et al.* 1995.

The fact that these bacteria are able to fix nitrogen is of great interest. Of no less importance is the fact that some of them, the obligate endophytes, are able to live in the interior of plants and colonise their tissues. This is done without apparently causing any adverse symptoms which would indicate a disease or a disturbance of the normal function of the plant.

Presently, the nitrogen fixing bacteria, which associate with plants, are of three categories: 1) rhizosphere, or bacteria that colonise the root surface; 2) facultative endophytes, which colonise the surface and the interior of the roots; 3) obligate endophytes, which colonise the root interior and the aerial tissues of the plants (Baldani *et al.*, 1997).

The gramineae are the plants most frequently associated with nitrogen fixing bacteria (Boddey R.M., 1987; James E.K. and Olivares F.L., 1997). To this family belongs maize, wheat, rice and sugarcane, along with several pasture grasses, all of them important food crops whose production consumes large quantities of industrial nitrogen fertilisers. All these plants show some association with endophytic nitrogen fixing bacteria and in laboratory experiments some show the ability to utilise the nitrogen fixed by the bacteria.

At the moment, there is no general evidence of endosymbiosis of the endophitic bacteria inside the plant cytoplasm. However, the fact that the

plant can use, under special conditions, nitrogen fixed by its associated bacteria is a positive indication that this system can be improved and adopted.

Recent work at the International Rice Research Institute (IRRI) in the Philippines and at Embrapa-Agrobiologia near Rio de Janeiro has shown that some varieties of wetland rice can obtain up to 20 or 25% of their N requirement from N2 fixation. Several species of N2-fixing bacteria have been found in the rhizosphere of this crop, but the roots and stems have also been found to be colonised by high numbers endophytic diazotrophs of the genera herbaspirillum and burkholderia.

In the case of sugar cane, in 15N studies carried out at Embrapa-Agrobiologia some cane varieties were found able to obtain over 60% of their N from N2 fixation (Urquiaga *et al.*, 1992) and these results have been confirmed by 15N natural abundance studies under field conditions. In this crop several species of endophytic diazotroph including acetobacter diazotrophicus, herbaspirillum and burkholderia have also been found in high numbers within plant tissues (Yoneyama *et al.*, 1997)

However, in both cases research has not yet revealed whether the observe N2 fixation is specifically associated with any one species of these bacteria, or whether more than one species work together to contribute to this activity. For this reason, as yet the association of the diazotrophs with these crops cannot be described as a "symbiosis" as the partners are not defined and no specific symbiotic structures have been revealed.

The role of BNF in Biofuel production

Trees and tropical grasses are regarded as the highest biomass yields per ha. In the case of sugar cane the sugar produced can be converted at low cost to ethanol and this makes it a most attractive source of biofuel.

In Brazil, cane breeding was conducted on low-N soils and without N fertiliser which evidently favoured BNF inputs such that Brazilian varieties can produce high cane yields with only inputs of N fertiliser. This is crucial to biofuel production as N fertiliser is produced from fossil fuels (natural gas), and if very large amounts are added to the crop, more fossil energy is required to produce the alcohol than is obtained from it when it is used as fuel (Boddey, 1995). Brazil manages to add 20% ethanol in all its gasoline and run approximately 3 million cars on pure alcohol from this source, while using only 4 million hectares of agricultural land (8% of Brazil's cropped area).

Another fact of interest is the versatility of the genotype in the gramineae family. Considering the family in terms of the number of individuals and the large area of the earth's surface which it dominates including

prairies, steppes, savannahs, pampas and paramos, it is evident that the genotype permits the gramineae to be one of the earth's most abundant families of vascular plants. Further search will, in all probability, reveal a strain which possesses the desired association and provide the key for solving the problem of an endosymbiosis, or its equivalent, between BFN and the gramineae plants.

The African oil palm will produce palm oil at a yield of approximately 4 tons/ha/year and the crop is generally found to have little response to N fertiliser addition. Recent studies at Embrapa-Agrobiologia have shown the presence of high numbers of endophytic diazotrophs within the fruits, stems and roots of these trees, which may explain their self sufficiency in nitrogen. Up to 20% of diesel oil can be substituted by palm oil without having to modify commercial diesel engines, and specially designed engines can run on pure palm oil, so this crop holds out a huge potential for the future production of biofuel.

As already mentioned, it uses little if any industrial nitrogen fertiliser; its by-products pollute much less; and the crop may have an important role in the forestation of a great portion of the deforested part of Amazonian region.

Of interest are the results of a research project co-ordinated by Pavan and Moreira Filho 1997 on nitrogen fixing bacteria in plants of different parts of Brazil.

We isolated from different plants over 40 strains that are now being studied in detail, which are associated with several families of epiphytic, parasitic and others plants that are part of the great biodiversity of the tropical and sub-tropical regions.

Our preliminary results based on their analysis of many plants belonging to different families, showed that all of the plants tested are associated with one or more bacteria able to fix nitrogen.

The type of association between the nitrogen fixing bacteria and plants is presently under study.

Acknowledgement

We would like to thank R.M. Bodley, I.J. Baldani, Elisabeth F. Pessoa and O. Frota-Pessoa for their help.

REFERENCES

Baldani, J.I., Caruso, L., Baldani, V.L.D., Goi S.R. and Döbereiner, J. (1997): *Soil Biol. Brochem.*, 29, n. 5, pp. 911-922.

Boddey, R.M. (1987): 'Methods for Quantification of Nitrogen Fixation Associated with Gramineae', *CCR Critical Reviews in Plant Sciences*, 6 (3), pp. 209-266.

Boddey, R.M. (1995): 'Biological Nitrogen Fixation in Sugar Cane: a Key to Energetically Viable Biofuel Production', *C.R.C. Crit. Rev. Plants Sci.*, 14, pp. 263-279.

Carpenter, S., Caraco, D.L., Howarth, R.W., Sharpey, A.N. and Smith, V.H. (1998): 'Nonpoint Pollution of Surface Waters with Phosphorus and Nitrogen Issues', *Ecology*, n. 3, Summer 1998. Published by the Ecological Society of America, Washington DC, esahq@esa.org, Available electronically http://esa.sesc.edu/

Döbereiner, J. and Day, J.M. (1976): 'Associative Symbioses in Tropical Grasses: Characterisation of Micro-organisms and Dinitrogen-Fixing Sites', in *Proceedings of the 1st International Symposium on Nitrogen fixation*, vol. 2, W.E. Newton and C.J. Nyman (ed.) (Washington State Univ. Press, Pullman wash), pp. 518-538.

Döbereiner, J., Baldani, V.L.D. and Baldani, J.I. (1995): *Como isolar e identificar bactérias diazotróficas de plantas não leguminosas*, Publicação EMBRAPA SPI: Itaguai, RJ. Embrapa CNPAB, 1995, 60 pp.

Dobzhansky, Th. (1962): *Mankind Evolving* (Bantam Books, USA).

Hardy, R.W.F. (1993): 'Biologiacal Nitrogen Fertilisation Present and Future Applications', pp. 109-117, in *Agriculture and Environmental Challenges*, J.P. Srivastava and H. Aldermam (eds.), *Proc. 13th Agric. Sector Symp.*, Washington DC, The World Bank.

Hardy, R.W.F. (1996): *Biological Nitrogen Fixation*, Nat. Res. Council Panel Chaired by Hardy, R.W.F., http://www.nap.edu/readingroom/book/bnf/preface.hteml

James, E.K. and Olivares, F.L. (1997): 'Infection and Colonisation of Sugar Cane and other Graminaceus Plants by Endophytic Diazotrophs', *Critical Reviews in Plant Sciences*, 17 (1), pp. 77-119.

Pavan, C. and Moreira Filho, C.A. (1998): 'Bacterias Fixadoras de Nitrogênio na Agricultura e na Biodiversidade', *Biotecnologia C&T*, n. 4, pp. 38-39, Brasil, www.biotecnologia.com.br

Raven, P.H. and Johnson, G.B. (1992): *Biology* (3rd edn., Mosby, Year Book Inc. USA).

Raven, P.H. (1998): *Biotechnology, Biodiversity, and Economic Development*, Biosafety Protocol Meeting Convention on Biological Diversity, Montreal, P. Q. Canada, August 19, 1998.

Urtiaga, S., Cruz, K.H.S. and Boddey, R.M. (1992): 'Contribution of Nitrogen Fixation to Sugarcane: Nitrogen 15 and Nitrogen Balance Estimates', *Soil Sciences Soc. of America Journal* 56, pp. 105-114.

Vitousek, P.M., Aber, J., Howart, R.W., Likens, G., Matson, P.A., Schindler, D.W., Schelinger, W.H. and Tilman, G.D. (1997): *Issues in Ecology*, n. 1, pp. 1-15, published by the Ecological Society of America, Washington DC, esahq@esa.org

Vitousek, *et al.* (1997): The same tittle above in a more complete publication. http://www.sda.edu/~ESA/

Yoneyama, T., Muraolca, T., Kim, T.H., Dacaney, E.V. and Nakanishi, Y. (1997): 'The Natural ^{16}N abundance of Sugarcane and neighbouring plants in Brazil, the Philippines and Miyako (Japan)', *Plant Soil*, 189, pp. 239-244.

PROSPECTS FOR GLOBAL SECURITY

RAJUL PANDYA-LORCH

Introduction

Although enough food is being produced today so that nobody should have to go hungry, about 840 million people are chronically undernourished, around 185 million pre-school children are seriously underweight for their age, and illnesses resulting from or exacerbated by hunger and malnutrition are widespread. As the world's population increases by an expected 80 million people every year over the next quarter century, assuring food security will be the central global challenge. Will there be enough food to meet the needs of current and future generations? And even if enough food is available, will all people have access to sufficient food to lead healthy and productive lives, that is, will they have the means to grow and/or purchase the needed food? Can, and will, global food security be attained or will food surpluses continue to co-exist with widespread hunger and malnutrition, further destabilizing and polarizing the world? What will it take to assure a world of food-secure people?

Following a brief discussion of food security concepts and an assessment of the current food security situation, this paper examines the outlook for global food security and identifies key actions required to assure global food security.

Food Security Concepts

The world is food secure when each and every person is assured of access at all times to the food required for a healthy and productive life. Food security is jointly determined by availability of food and access to food. Availability of food does not guarantee access to food, but access to food is contingent on there being food available (von Braun *et al.*, 1992). National, regional, or local availability of food is a function of food pro-

duction, stockholding, and trade. National access to food from international markets is determined by world food prices and foreign exchange availability. Household availability of food requires that food be available at local or regional markets, which is determined by market operations, infrastructure, and information flows. Access to food by households and individuals is usually conditioned by income: the poor commonly lack adequate means to secure their access to food.

Food security at any level does not guarantee food security at any other level (von Braun *et al.*, 1992). For example, household food security does not necessarily mean that all individuals in that household have access to the needed food. Some members of a household may be denied their full share of the needed food. Intra-household inequality in distribution of food, with women in particular eating less than their share of household food, is observed quite often. Similarly, regional or national food security does not necessarily lead to household or individual food security; the available food may not be distributed according to needs and households or individuals may not have equitable access to it. And, of course, global food availability does not mean universal food security. There may be marked national, regional, household, and individual differences in access to food.

Current World Food Security Situation

Despite impressive food production growth in recent decades such that enough food is available to meet the basic needs of each and every person in the world, not all people are food secure. If available food were evenly distributed, each person would be assured of 2,700 calories per day, 20 percent more than in 1961-63 (FAO, 1997a). However, available food is neither evenly distributed nor fully consumed among or within countries. Forty-two countries were unable to assure minimum requirements of 2,200 calories per person per day for their populations during 1992-94, even if available food had been evenly distributed within each country (FAO, 1997a). Of these countries, 29 were in Africa, 6 in Asia, 3 in Latin America and the Caribbean, 3 in Eastern Europe and the former Soviet Union, and 1 in the Middle East.

In the developing world as a whole, about 840 million people – 20 percent of the population – were chronically undernourished during 1990-92, lacking economic or physical access to sufficient food to lead healthy and productive lives (FAO, 1996a). East Asia was home to 32 percent of the world's undernourished, South Asia to 30 percent, and Sub-Saharan Africa to 26 percent. China and India together accounted for 45 percent of the world's undernourished people (FAO, 1996d). Progress is being made in reducing the magnitude and prevalence of undernourished people. There

were about 80 million fewer undernourished people in 1990-92 relative to 1969-71, while an additional 1.5 billion people were being adequately fed. The share of undernourished people in the population declined in more than 55 countries between 1969-71 and 1990-92 (FAO, 1996d), contributing to a reduction in the share of undernourished people in the developing world's population from 35 to 21 percent during this period. Most of the improvements in food security have taken place in East Asia where the number of undernourished people fell from 475 million in 1969-71 to 268 million in 1990-92. Nevertheless, with two-thirds of the developing world's undernourished, South and East Asia remain key areas of food security concern. And a new "flash-point" or locus of hunger and food insecurity has emerged in Sub-Saharan Africa, where the number of undernourished people doubled between 1969-71 and 1990-92 to 215 million, and the proportion of the population that is undernourished rose from 38 to 43 percent.

Child malnutrition is another indicator of food insecurity. The number of malnourished children rose during the 1980s from 164 million to 184 million, although due to population growth their share of the preschool children population declined slightly from 37.8 percent to 34.3 percent (UN ACC/SCN, 1992). One-third of all pre-school children are still underweight in the developing world. About 101 million underweight children are in South Asia, 44 million in East Asia and 28 million in Sub-Saharan Africa. About 60 percent of the preschool children in South Asia are underweight compared to 30 percent in Sub-Saharan Africa and East Asia respectively and 8 percent in Latin America and the Caribbean.

Micronutrient deficiencies are also widespread in the developing world, even where caloric consumption is adequate. Micronutrient deficiencies have detrimental effects on human health and productivity. About 2 billion people are affected by iron deficiency, around 1.6 billion people are at risk of iodine deficiency, and 40 million children suffer from Vitamin A deficiency (FAO, 1996b).

In addition to those who are already food-insecure and show symptoms or consequences of food insecurity, there are many others worldwide who live with the risk of food insecurity: their incomes are so low that any sudden shock such as loss of employment or price fluctuations could tip them into food insecurity. These vulnerable people must also be taken into account when considering the world food security situation.

Earlier, it was noted that food security is jointly determined by availability of food and access to food. With regard to availability of food, food production growth in recent decades has been impressive. Between 1961-63 and 1994-96, food production increased by 119 percent worldwide while it increased 200 percent in developing countries as a group, with particularly large increases in the developing countries of Asia. Even in the developing

countries of Africa, where concerns regarding food security are greatest, food production increased by 120 percent during this period. Between 1961-63 and 1994-96, cereal production worldwide more than doubled to 1.97 billion tons and almost tripled in developing countries to 1.14 billion tons; meat production almost tripled worldwide to 208 million tons and quintupled in developing countries to 107 million tons; and production of roots and tubers doubled in developing countries increased to 436 million tons.

Worldwide, food production growth more than kept pace with population growth; *per capita* food production increased by 40 percent between 1961-63 and 1994-96. In the developing countries as a group, *per capita* food production increased 47 percent during this period. However, food production performance varied widely among developing regions; while *per capita* food production increased 67 percent in the developing countries of Asia, less food was produced per person in the developing countries of Africa in the mid-1990s than in the beginning of the 1960s. Between 1961-63 and 1994-96, cereal production per person worldwide increased by 20 percent to 350 kilograms while it increased by 28 percent to 252 kilograms in the developing world; and meat production per person worldwide increased by 55 percent to 37 kilograms while it increased by 242 percent to 24 kilograms in the developing world (FAO, 1997a).

There are indications that growth in food production has begun to lag in recent years. The annual rate of growth of global cereal production dropped from 2.6 percent during 1967-82 to 1.3 percent during 1982-94, while the annual rate of increase in cereal yields slowed from 2.3 percent to 1.5 percent between these two periods (Rosegrant *et al.*, 1997). After steadily increasing during the 1960s and 1970s, world grain production per person has fallen by about 1 percent annually over the past decade (Brown *et al.*, 1995). Yields of rice and wheat have been constant over the past few years in Asia, which is a significant producer (Pinstrup-Andersen, 1994). It is becoming increasingly difficult to maintain the yield gains already achieved, let alone to increase yields, in the high-potential or more-favored areas, while in the less-favored areas, which are home to many of the world's food-insecure people, yields are low and variable (Hazell, 1995).

Outlook for Global Food Security[1]

Projections of food production and consumption to the year 2020 offer some signs of progress, but prospects of a food-secure world – a world in

[1] This section draws upon Pinstrup-Andersen *et al.* (1997).

which each and every person is assured of access at all times to the food required to lead a healthy and productive life -- remain bleak if the global community continues with business as usual. IFPRI's global model, the International Model for Policy Analysis of Commodities and Trade (IMPACT),[2] projects that 150 million children under the age of six years will be malnourished in 2020, just 20 percent fewer than in 1993 (fig. 1).[3] One out of every four children will be malnourished in 2020, down from 33 percent in 1993. Child malnutrition is expected to decline in all major developing regions except Sub-Saharan Africa, where the number of malnourished children could increase by 45 percent between 1993 and 2020 to reach 40 million. In South Asia, home to half of the world's malnourished children in 1993, the number of malnourished children is projected to decline by more than 30 million between 1993 and 2020, but the incidence of malnutrition is so high that, even with this reduction, two out of five children could remain malnourished in 2020 (fig. 2). With more than 70 percent of the world's malnourished children, Sub-Saharan Africa and South Asia are expected to remain "hot spots" of child malnutrition in 2020.

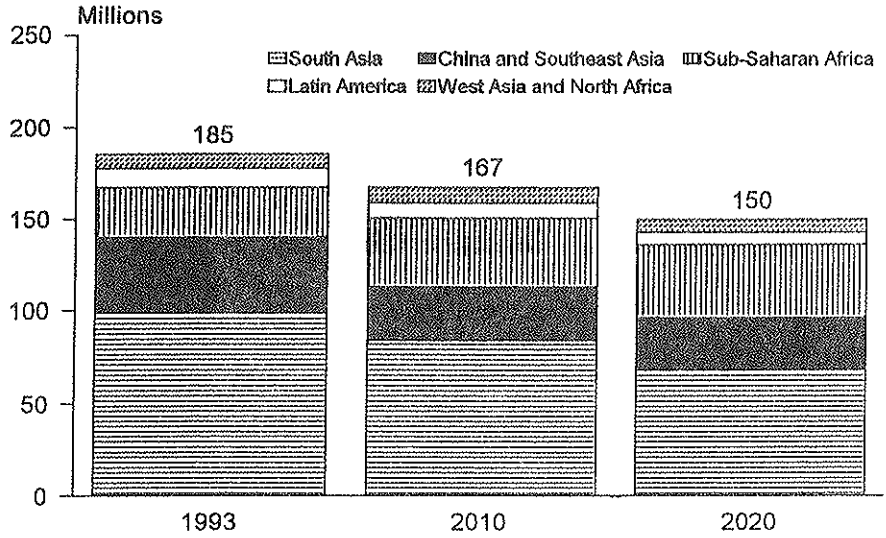
Projections by FAO on the number of food-insecure people paint a similarly mixed picture.[4] FAO projects that 680 million people, 12 percent of the developing world's population, could be food insecure in 2010, down from 840 million in 1990-92 (fig. 3). Food insecurity is expected to diminish rapidly in East Asia and, to a lesser extent, in South Asia and Latin America, but it could accelerate substantially in Sub-Saharan Africa and West Asia and North Africa. Sub-Saharan Africa and South Asia, home to a projected 70 percent of the world's food-insecure people in 2010, will be the locus of hunger in the developing world. In fact, Sub-Saharan Africa's share of the world's food-insecure population is projected to almost quadruple between 1969-71 and 2010 from 11 to 39 percent (FAO, 1996c). By 2010, every third person in Sub-Saharan Africa is likely to be food insecure compared with every eighth person in South Asia and every twentieth person in East Asia. These disturbing figures reflect widespread poverty and poor health.

---

[2] IMPACT covers 37 countries and regions (which account for virtually all of the world's food production and consumption) and 17 commodities (including all cereals, soybeans, roots and tubers, meats, and dairy products) (Rosegrant *et al.*, 1995; Rosegrant *et al.*, 1997).

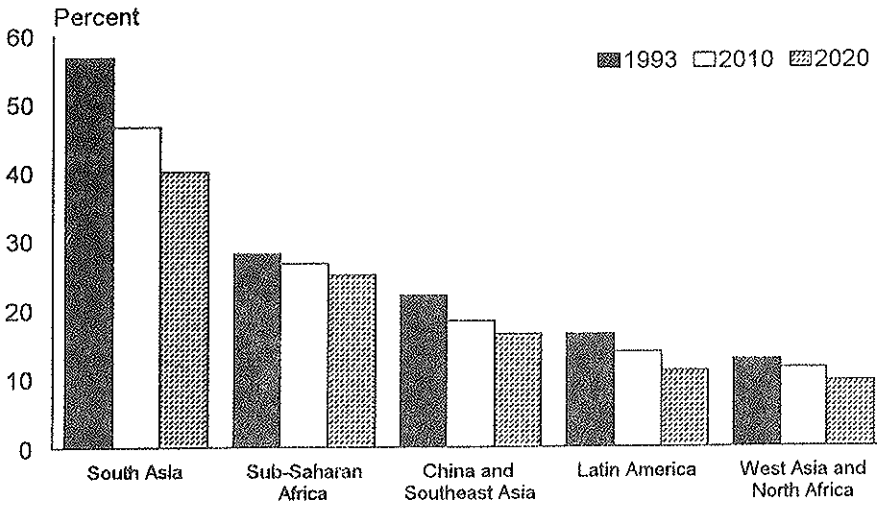
[3] Malnourished children are those whose weight-for-age is more than two standard deviations below the weight-for-age standard set by the U.S. National Center for Health Statistics and adopted by many United Nations agencies in assessing the nutritional status of persons in developing countries.

[4] FAO classifies these people as chronically undernourished; that is, their access to per capita food supplies is less than 1.55 times the basal metabolic rate.



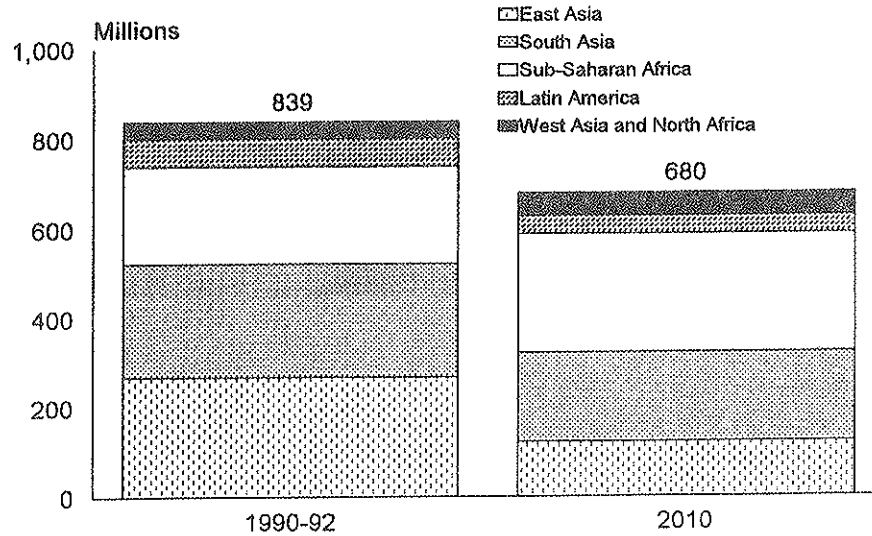
Source: IFPRI IMPACT simulations.

Fig. 1. Number of malnourished children, 1993, 2010, and 2020.



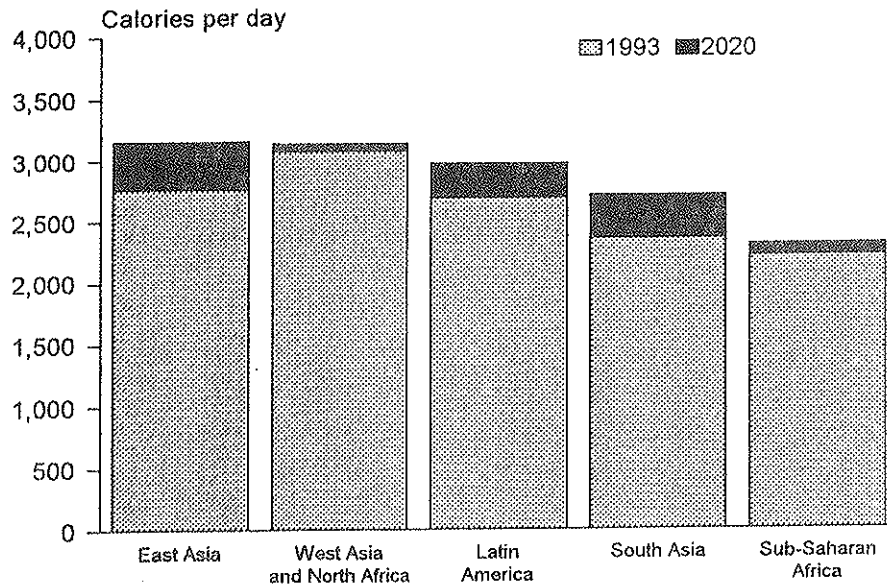
Source: IFPRI IMPACT simulations.

Fig. 2. Percentage of malnourished children, 1993, 2010, and 2020.



Source: Food and Agriculture Organization of the United Nations, *Food, Agriculture, and Food Security: Developments since the World Food Conference and Prospects* World Food Summit Technical Background Document 1 (Rome, 1996).

Fig. 3. Number of food-insecure people, 1990-92 and 2010.



Source: IFPRI IMPACT simulations.

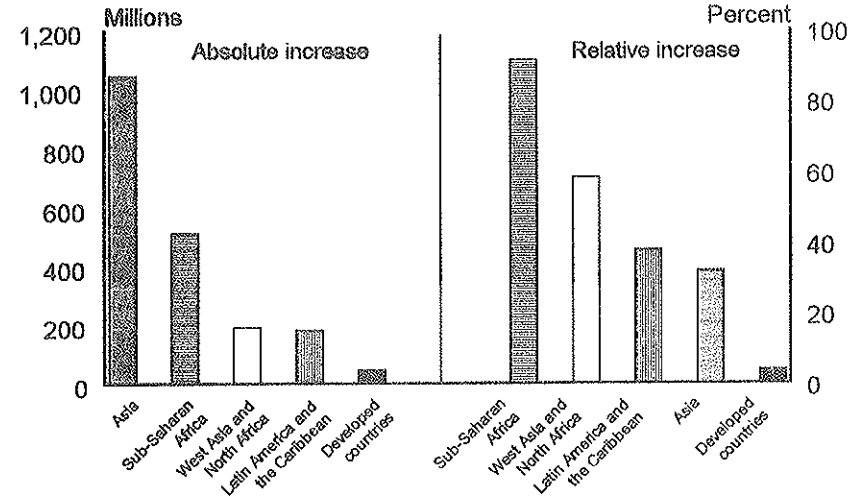
Fig. 4. Daily per capita calorie availability, 1993 and 2020.

Worldwide, *per capita* availability of food is projected to increase around 7 percent between 1993 and 2020, from about 2,700 calories per person per day in 1993 to about 2,900 calories. Increases in average *per capita* food availability are expected in all major regions. China and East Asia are projected to experience the largest increase, and West Asia and North Africa the smallest (fig. 4). The projected average availability of about 2,300 calories per person per day in Sub-Saharan Africa is just barely above the minimum required for a healthy and productive life. Since available food is not equally distributed to all, a large proportion of the region's population is likely to have access to less food than needed.

Related to this is an increasing gap between food demand and production in several parts of the world. Demand for food is influenced by a number of forces, including population growth and movements, income levels and economic growth, human resource development, and lifestyles and preferences. In the next several decades, population growth will contribute to increased demand for food. The United Nations recently scaled back its population projections, but even with these reduced estimates, almost 80 million people are likely to be added to the world's population each year during the next quarter century, increasing world population by 35 percent from 5.69 billion in 1995 to 7.67 billion by 2020 (UN, 1996). More than 95 percent of the population increase is expected in developing countries, whose share of global population is projected to increase by 79 percent in 1995 to 84 percent in 2020. Over this period, the absolute population increase will be highest in Asia, but the relative increase will be greatest in Sub-Saharan Africa, where the population is expected to almost double by 2020 (fig. 5).

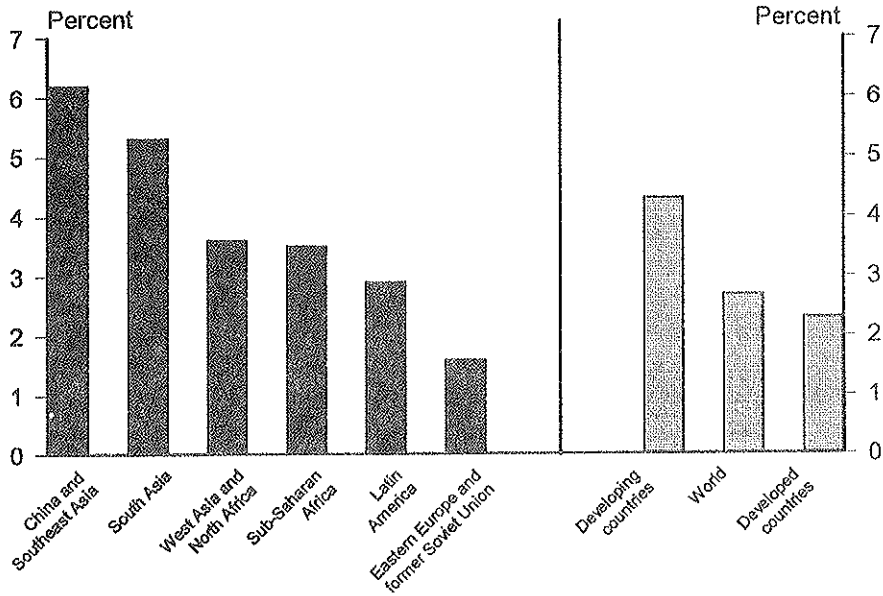
At the same time, urbanization will contribute to changes in the types of food demanded. Much of the population increase in developing countries is expected in the cities; the developing world's urban population is projected to double over the next quarter century to 3.6 billion (UN, 1995). Urbanization profoundly affects dietary and food demand patterns: the increasing opportunity cost of women's time, changes in food preferences caused by changing lifestyles, and changes in relative prices associated with rural-urban migration lead to more diversified diets with shifts from basic staples such as sorghum, millet, and maize to other cereals such as rice and wheat that require less preparation and to milk and livestock products, fruits and vegetables, and processed foods.

People's access to food depends on income. Currently, more than 1.3 billion people are absolutely poor, with incomes of a dollar a day or less per person, while another 2 billion people are only marginally better off (World Bank, 1997a). Income growth rates have varied considerably between



Source: United Nations, *World Population Prospects: The 1996 Revisions* (New York, 1996).
Note: Medium-variant projections.

Fig. 5. Absolute and relative population increases, 1995-2020.



Source: IFPRI IMPACT simulations.

Fig. 6. Projected average annual income growth rates, 1993-2020.

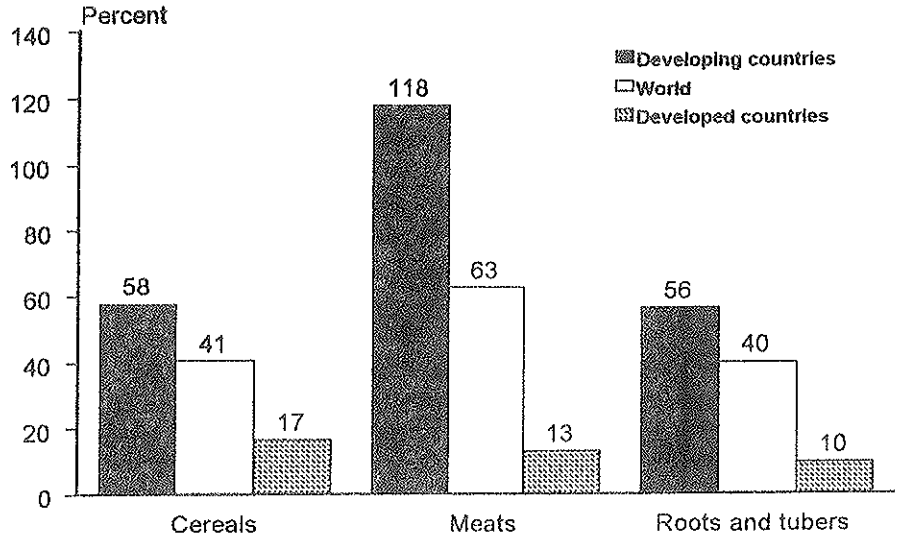
regions in recent years, with Sub-Saharan Africa and West Asia and North Africa struggling with negative growth rates while East Asia was experiencing annual growth rates exceeding 7 percent (World Bank, 1997b). Prospects for economic growth during the next quarter century appear favorable, with global income growth projected to average 2.7 percent per year between 1993 and 2020 (fig. 6). The projected income growth rates for developing countries as a group are almost double those for developed countries. Growth rates are projected to be lowest in Eastern Europe and the former Soviet Union. Even Sub-Saharan Africa is expected to experience positive *per capita* income growth between 1993 and 2020, although it will be quite low. However, unless significant and fundamental changes occur in many developing countries, disparities in income levels and growth rates both between and within countries are likely to persist, and poverty is likely to remain entrenched in South Asia and Latin America and to increase considerably in Sub-Saharan Africa.

IFPRI projects global demand for cereals to increase by 41 percent between 1993 and 2020 to reach 2,490 million metric tons, for meat demand to increase by 63 percent to 306 million tons, and for roots and tubers demand to increase by 40 percent to 855 million tons (fig. 7).[5] Most of the increases in demand between 1993 and 2020 are projected to occur in developing countries, which will account for more than 80 percent of the increase in global cereal demand, nearly 90 percent of the increase in meat demand, and more than 90 percent of the increase in demand for roots and tubers. Among the major developing regions, Sub-Saharan Africa is expected to experience the largest percentage increase in demand for all the major food commodities, albeit from low levels (fig. 8).

Demand for cereals for feeding livestock will increase considerably in importance in coming decades, especially in developing countries, in response to strong demand for livestock products. Between 1993 and 2020, developing countries' demand for cereals for animal feed is projected to double while demand for cereals for food for direct human consumption is projected to increase by 47 percent (fig. 9). By 2020, 24 percent of the cereal demand in developing countries will be for feed, compared with 19 percent in 1993. However, in absolute terms, the increase in cereal demand for food will be higher than for feed. In developed countries, the increase in cereal demand for feed will outstrip the increase in cereal demand for food in both absolute and relative terms.

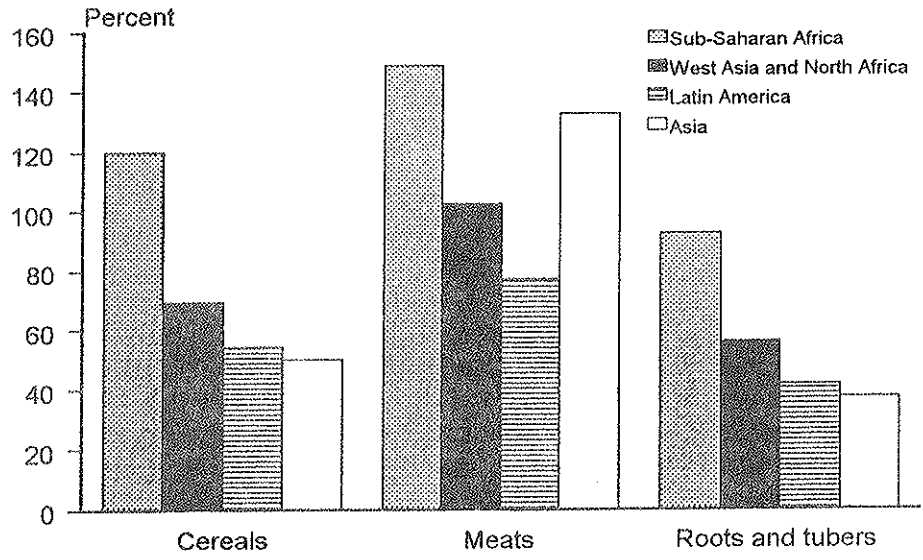
Because of substantial increases in demand for livestock products,

[5] All tons in this paper are metric tons.



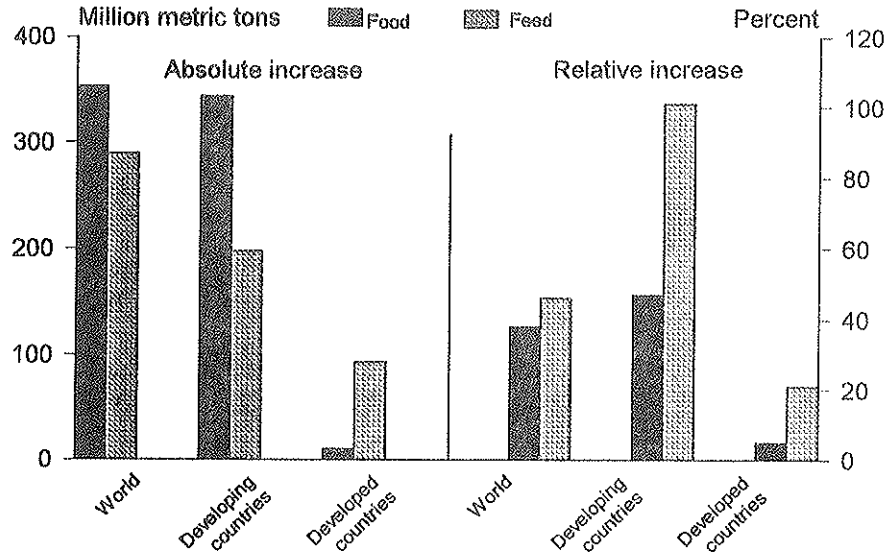
Source: IFPRI IMPACT simulations.

Fig. 7. Increase in total demand for cereals, meats, and roots and tubers, 1993-2020.



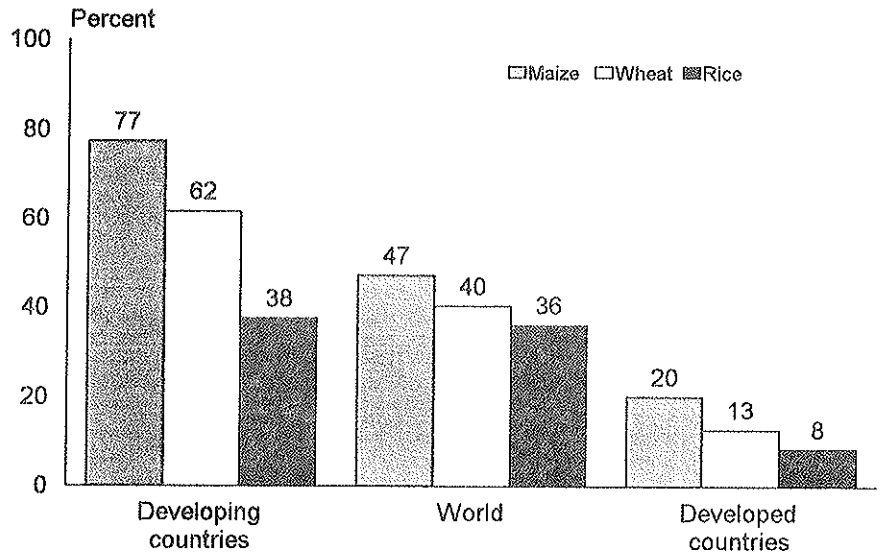
Source: IFPRI IMPACT simulations.

Fig. 8. Increase in total demand for cereals, meats, and roots and tubers in major developing regions, 1993-2020.



Source: IFPRI IMPACT simulations.

Fig. 9. Absolute and relative increase in food and feed demand for cereals, 1993-2020.



Source: IFPRI IMPACT simulations.

Fig. 10. Increase in total demand for major cereal commodities, 1993-2020.

especially in developing countries where primarily maize and other coarse grains are used for animal feed, demand for maize is projected to increase faster than for other cereals in both developed and developing countries (fig. 10). Global demand for maize is projected to grow at an annual rate of 1.4 percent between 1993 and 2020, followed by wheat at 1.3 percent and rice at 1.2 percent. In China and India, for instance, demand for maize and other grains for feed is projected to increase by around 3 percent per year between 1993 and 2020.

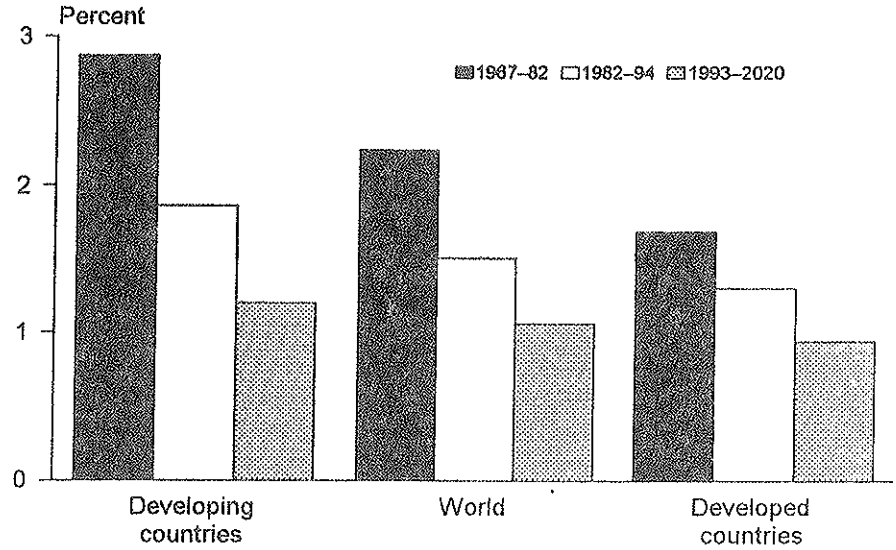
How will the expected increases in cereal demand be met? Not by expansion in cultivated area. IMPACT projections indicate that the area under cereals will increase by only 5.5 percent or 39 million hectares between 1993 and 2020, almost two-thirds of which will be in Sub-Saharan Africa. Since growth in cultivated area is unlikely to contribute much to future production growth, the burden of meeting increased demand for cereal rests on improvements in crop yields. However, the annual increase in yields of the major cereals is projected to slow down during 1993-2020 in both developed and developing countries (fig. 11). This is worrisome given that yield growth rates were already on the decline. Two of the key reasons for slow cereal yield growth rates are:

1. In regions where input use is high, such as Asia, farmers are approaching economically optimum yield levels, making it more difficult to sustain the same rates of yield gains.

2. Declining world cereal prices are causing farmers to switch from cereals to other, more profitable crops and are causing governments to slow their investment in agricultural research and irrigation and other infrastructure.

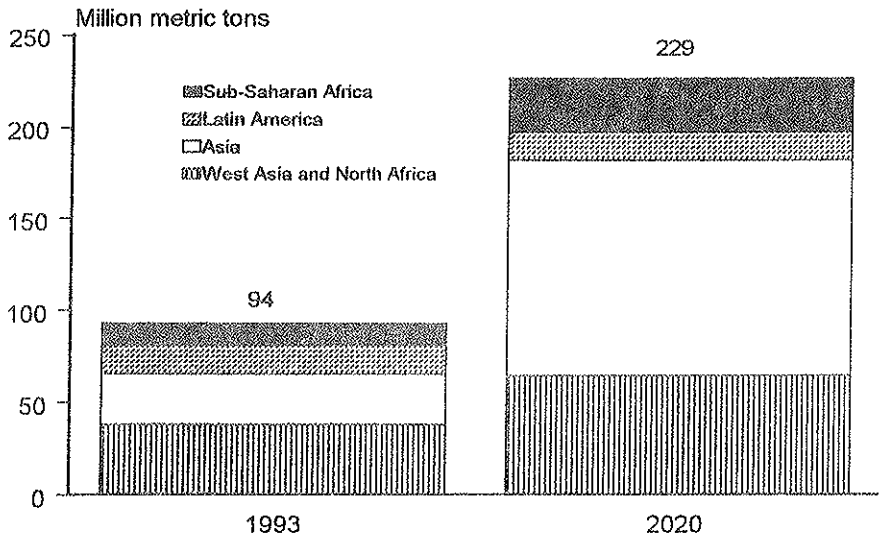
With the projected slowdowns in area expansion and yield growth, cereal production in developing countries as a group is also forecast to slow to an annual rate of 1.5 percent during 1993-2020 compared with 2.3 percent during 1982-94. This figure is still higher, however, than the 1.0 percent annual rate of growth projected for developed countries during 1993-2020.

Cereal production in developing countries will be insufficient to meet the expected increase in demand. As a group, developing countries are projected to more than double their net imports of cereals (the difference between demand and production) between 1993 and 2020 (fig. 12). With the exception of Latin America, all major developing regions are projected to increase their net cereal imports: the quadrupling of Asia's net imports will be driven primarily by rapid income growth, while the 150 percent increase forecast for Sub-Saharan Africa will be driven primarily by its continued poor performance in food production. While wheat is expected to constitute the bulk of the developing world's net cereal imports in 2020, the share of maize is forecast to sharply increase from 19 percent in 1993 to 27



Source: IFPRI IMPACT simulations.

Fig. 11. Annual growth in cereal yields, 1967-82, 1982-94, and 1993-2020.



Source: IFPRI IMPACT simulations.

Fig. 12. Net cereal imports of major developing regions, 1993 and 2020.

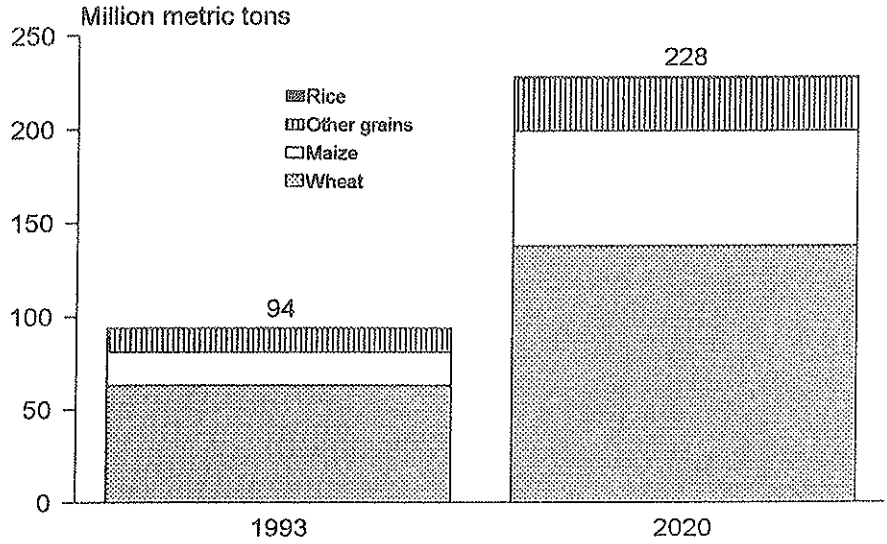
percent primarily because of the rapid increase in demand for meat (fig. 13). Trade in rice is forecast to remain negligible.

With continued population growth, rapid income growth, and changes in lifestyles, demand for meat is expected to rise rapidly in developing countries. IMPACT projections indicate that total demand for meat will increase by 2.9 percent per year during 1993-2020 in developing countries and by 0.5 percent per year in developed countries. Worldwide, demand for meat is projected to increase by 1.8 percent per year, with demand for poultry expected to increase fastest at an annual rate of 2.1 percent, compared with 1.5 percent for beef. In *per capita* terms, demand for meat products is projected to increase by almost 50 percent in developing countries to 31 kilograms in 2020, and by 4 percent in developed countries to 81 kilograms. In 1993, developing countries accounted for 47 percent of world meat demand; by 2020, they are projected to account for 63 percent. Meat production is expected to grow by 2.7 percent per year in developing countries during 1993-2020 (compared with 5.9 percent during 1982-94) and by 0.8 percent in developed countries (compared with 0.9 percent during 1982-94). Despite high rates of production growth, developing countries as a group are projected to increase their net meat imports 20-fold, reaching 11.5 million tons in 2020 (fig. 14). Latin America will continue to be a net exporter of meat, but Asia will switch from being a small net exporter to a large net importer. Beef is expected to constitute 46 percent of the developing world's net meat imports in 2020, poultry 30 percent, pigmeat 13 percent, and sheep and goatmeat 11 percent.

Projections of future fish consumption are scarce. FAO projections suggest that direct human consumption of fish will increase from 75-80 million tons in 1994/95 to 110-120 million tons in 2010 (Delgado and Courbois, 1997). Much of the increase in fish consumption is projected to occur in East Asia and, to a lesser extent, in North America and Australia. China's per capita consumption of fish is predicted to double from 9.8 kilograms in 1990 to 20 kilograms in 2010, driven primarily by income increases, and will be met increasingly from aquaculture production.

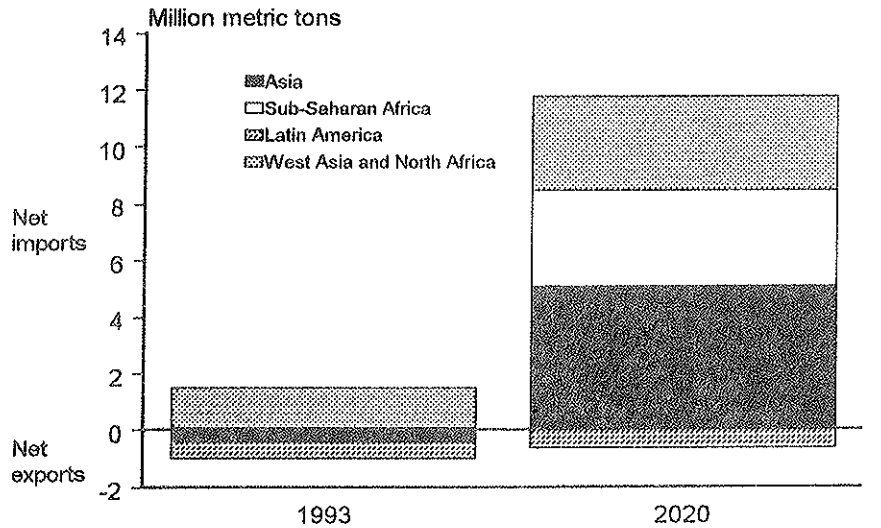
With sustainable production from the world's natural fish stock at its limit, supplies from capture fisheries have stabilized at around 90 million tons after nearly three decades of steady growth. Aquaculture, however, has become the fastest-growing food production system in the world, with global production increasing on average by more than 11 percent annually between 1990 and 1995 (Ahmed, 1997). The share of global fish production contributed by aquaculture rose from 13 percent to 19 percent in the period 1990-95 (FAO, 1997b).

Real fish prices have remained relatively stable since 1970, while real



Source: IFPRI IMPACT simulations.

Fig. 13. Composition of net cereal imports by developing countries, 1993 and 2020.



Source: IFPRI IMPACT simulations.

Fig. 14. Net trade in meat by major developing regions, 1993 and 2020.

beef prices have declined substantially and are now less than one-third of the 1970 price. Some researchers report an emerging consensus that real fish prices are likely to rise by about 10 percent by 2020 (Delgado and Courbois, 1997), while IMPACT projections suggest that beef prices will decline by about 5 percent between 1993 and 2020, implying a long-run increase in the relative fish-beef price and therefore major adjustments in the world markets for both fish and beef.

Net imports are a reflection of the gap between production and market demand. For many of the poor, the gap between food production and human needs is likely to be even wider than that between production and demand, because many of these people are priced out of the market, even at low food prices, and are unable to exercise their demand for needed food. The higher-income developing countries, notably those of East Asia, will be able to fill the gap between production and demand through commercial imports, but the poorer countries may be forced to allocate foreign exchange to other uses and thus might not be able to import food in needed quantities. It is the latter group of countries, including most of those in Sub-Saharan Africa and some in Asia, that will remain a challenge and require special assistance to avert widespread hunger and malnutrition.

Required Action

The action required to assure a food-secure world is known. Much thought and effort have been expended to identify priority action at the individual, household, community, national, regional, and global levels. At the World Food Summit convened by the Food and Agriculture Organization of the United Nations (FAO) in November 1996, leaders from around the world signed the Rome Declaration on World Food Security, reaffirming "the right of every person to have access to safe and nutritious food, consistent with the right to adequate food and the fundamental right of everyone to be free from hunger" (FAO, 1996e). They pledged "[their] political will and [their] common and national commitment to achieving food security for all and to an ongoing effort to eradicate hunger in all countries, with an immediate view to reducing the number of undernourished people to half their present level no later than 2015" (FAO, 1996e). Toward this end, they made seven commitments:

— Ensure an enabling political, social, and economic environment designed to create the best conditions for the eradication of poverty and for durable peace, based on full and equal participation of women and men, which is most conducive to achieving sustainable food security for all.

— Implement policies aimed at eradicating poverty and inequality and improving physical and economic access by all, at all times, to sufficient, nutritionally adequate and safe food and its effective utilization.

— Pursue participatory and sustainable food, agriculture, fisheries, forestry, and rural development policies and practices in high and low potential areas, which are essential to adequate and reliable food supplies at the household, national, regional, and global levels, and combat pests, drought, and desertification, considering the multifunctional character of agriculture.

— Strive to ensure that food, agricultural trade, and overall trade policies are conducive to fostering food security for all through a fair and market-oriented world trade system.

— Endeavor to prevent and be prepared for natural disasters and man-made emergencies and to meet transitory and emergency food requirements in ways that encourage recovery, rehabilitation, development, and a capacity to satisfy future needs.

— Promote optimal allocation and use of public and private investments to foster human resources, sustainable food, agriculture, fisheries and forestry systems, and rural development, in high and low potential areas; and

— Implement, monitor, and follow-up this Plan of Action at all levels in cooperation with the international community (FAO, 1996e).

A detailed plan of action seeks to achieve the goals included in these seven commitments.

The International Food Policy Research Institute (IFPRI), in its initiative on "A 2020 Vision for Food, Agriculture, and the Environment", has developed the 2020 Vision of "a world where every person has access to sufficient food to sustain a healthy and productive life, where malnutrition is absent, and where food originates from efficient, effective, and low-cost food systems that are compatible with sustainable use of natural resources" (IFPRI, 1995). Sustained action is required in six priority areas to realize the 2020 Vision:

— Strengthen the capacity of developing-country governments to perform appropriate functions, such as maintaining law and order, establishing and enforcing property rights, promoting and assuring private-sector competition in markets, and maintaining appropriate macroeconomic environments. Predictability, transparency, and continuity in policymaking and enforcement must be assured. The efforts of the past decade to weaken developing-country governments must be turned around. More effective local and national governments are essential for other partners, such as individuals, households, communities, nongovernmental organizations (NGOs),

and the private sector, to contribute to food security. Governments should facilitate food security for all households and individuals, not by physically delivering needed foods to all citizens but by facilitating a social and economic environment that provides all citizens with the opportunity to assure their food security.

— Invest more in poor people in order to enhance their productivity, health, and nutrition and to increase their access to remunerative employment and productive assets. Governments, local communities, and NGOs should assure access to and support for a complete primary education for all children, with immediate emphasis on enhancing access by female and rural children; assure access to primary health care, including reproductive health services, for all people; improve access to clean water and sanitation services; provide training for skill development in adults; and strengthen and enforce legislation and provide incentives for empowerment of women to gain gender equality. Improved access by the rural poor, especially women, to productive resources can be facilitated through land reform and sound property rights legislation, strengthened credit and savings institutions, more effective rural labor markets, and infrastructure for small-scale enterprises. Social safety nets for the rural poor are urgently needed. Direct transfer programs, including programs for poverty relief, food security, and nutrition intervention, are needed in many countries at least in the short term and must be better targeted to the poor. Efforts must be made to lower fertility rates and slow population increases. Strategies to reduce population growth rates include providing full access to reproductive health services to meet unmet demand for contraception; eliminating risk factors that promote high fertility, such as high rates of infant mortality or lack of security for women who are dependent on their children for support because they lack access to income, credit, or assets; and providing young women with education. Female education is one of the most important investments for assuring food security.

— Accelerate agricultural productivity by strengthening agricultural research and extension systems in and for developing countries. The key role of the agriculture sector in meeting food needs and fostering broad-based economic growth and development must be recognized and exploited. To make this happen, agricultural research systems must be mobilized to develop improved agricultural technologies, and extension systems must be strengthened to disseminate improved technologies. Investments in strategic international and regional agricultural research with large potential international benefits should be expanded to better support national efforts. Biotechnology research in national and international

research systems should be expanded to support sustainable intensification of small-scale agriculture in developing countries. Effective partnerships between developing-country research systems, international research institutions, and private- and public-sector research institutions in industrialized countries should be forged to assure relevance of research and appropriate distribution of responsibilities and to bring biotechnology to bear on the agricultural problems of developing countries. Developing countries can address funding and personnel constraints by providing incentives to the private sector to engage in such research, by collaborating with international research programs, and by seeking private- and public-sector partners in industrialized countries. They should be encouraged to adopt regulations that provide an effective measure of biosafety without crippling the transfer of new products to small farmers.

— Promote sustainable agricultural intensification and assure sound management of natural resources. Public- and private-sector investments in infrastructure, market development, natural resource conservation, soil improvements, primary education and health care, and agricultural research must be expanded in areas with significant agricultural potential, fragile soils, and large concentrations of poverty to effectively address their problems of poverty, food insecurity, and natural resource degradation before they worsen or spill over into other regions. In areas of low current productivity but significant agricultural potential, public policy and public-sector investment should promote sustainable use of existing natural resources to enhance the productivity of agriculture and other rural enterprises. Incentives should be provided to farmers and local communities to invest in, and protect, natural resources and to restore degraded lands. Clearly specified systems of rights to use and manage natural resources, including land, water, and forests, should be established and enforced. Local control over natural resources must be strengthened, and local capacity for organization and management improved. Farmers and communities should be encouraged to implement integrated soil fertility programs in areas with low soil fertility through policies to assure long-term property rights to land, access to credit, improved crop varieties, and information about production systems; through effective and efficient markets for plant nutrients, and investments in infrastructure and transportation systems; and through temporary fertilizer subsidies where prices are high due to inadequate infrastructure or poorly functioning markets. Integrated pest management programs should be promoted as the central pest management strategy to reduce use of chemical pesticides, remove pesticide subsidies, and increase farmer participation in developing effective and appropriate strate-

gies of pest management. Water policies should be reformed to make better use of existing water supplies by providing appropriate incentives to water users, improving procedures for water allocation, and developing and disseminating improved technology for water supply and delivery.

— Develop effective, efficient, and low-cost agricultural input and output markets. To obtain gains from improved efficiency and reduced costs of marketing agricultural inputs and outputs, governments should phase out inefficient state-run firms in agricultural input and output markets and create an environment conducive to effective competition among private agents in order to provide efficient and effective services to producers and consumers. Governments should identify their role in agricultural input and output markets and strengthen their capacity to perform this role better while disengaging itself from functions that should be undertaken by the private sector. Policies and institutions that favor large-scale, capital-intensive enterprises over small-scale, labor-intensive ones should be removed. Market infrastructure of a public-goods nature, such as roads, electricity, and communications facilities, should be developed and maintained by direct public-sector investment or effective regulation of private-sector investment. Governments should develop and enforce standards, weights and measures, and regulatory instruments essential for effective functioning of markets. Development of small-scale credit and savings institutions should be facilitated. Technical assistance and training could be provided to create or strengthen small-scale, labor-intensive competitive rural enterprises in trade, processing, and related marketing activities.

— Expand and realign international assistance and improve its efficiency and effectiveness. The current downward trend in international development assistance must be reversed, and industrialized countries allocating less than the United Nations target of 0.7 percent of their gross national product (GDP) should rapidly move to that target. Official development assistance, which is only a small fraction of the resources required by developing countries, must be allocated to effectively complement national and local efforts. Official government-to-government assistance should be made available primarily to countries that have demonstrated commitment to reducing poverty, hunger, and malnutrition and to protecting the environment. International development assistance must be realigned to low-income developing countries, primarily in Sub-Saharan Africa and South Asia where the potential for further deterioration of food security and degradation of natural resources is considerable. In higher-income developing countries, concessional aid such as grants should be replaced by internationally available commercial capital, freeing resources

for the low-income countries. To improve effectiveness of aid, each recipient country should develop a coherent strategy for achieving its goals related to food security, poverty, and natural resources, and should identify the most appropriate uses of international assistance.

Conclusions

Food insecurity has long been perceived by some to be primarily a problem of insufficient food production rather than insufficient access to food. Yet, as enough food is being produced to meet the basic needs of every person in the world, it is evident that the persistence of food insecurity – about 840 million chronically undernourished people and 185 million malnourished children – is increasingly attributable to difficulties in accessing sufficient food. Food-insecure people simply do not have the means to grow and/or purchase the needed food. Empowering every individual to have access to remunerative employment, to productive assets such as land and capital, and to productivity-enhancing resources such as appropriate technology, credit, education, and health care is essential. Besides enabling every person to acquire the means to grow and/or purchase sufficient food to lead healthy and productive lives, assuring a food-secure world calls for producing enough food to meet increasing and changing food needs and for meeting food needs from better management of natural resources.

With foresight and decisive action, we can create the conditions that permit food security for all people in coming years. The action required is not new or unknown; for instance, we know that increased productivity in agricultural production helps not only to produce more food at lower unit costs and make more efficient use of resources but also to raise the incomes of farmers and others linked to agriculture and thus improve their capacity to purchase needed food. The action program outlined earlier will require all relevant parties – individuals, households, farmers, local communities, the private sector, civil society, national governments, and the international community – to work together in new or strengthened partnerships; it will require a change in behavior, priorities, and policies; and it will require strengthened cooperation between developing and industrialized countries and among developing countries. The world's natural resources are capable of supporting sustainable food security for all people, if current rates of degradation are reduced and replaced by appropriate technological change and sustainable use of natural resources.

We have the means to assure a food-secure world; let us act to make it a reality for each and every person.

REFERENCES

Ahmed, M. (1997): 'Policy Issues Deriving from the Scope, Determinants of Growth, and Changing Structure of Supply of Fish and Fishery Products in Developing Countries', paper presented at the International Consultation on Fisheries Policy Research in Developing Countries: Issues, Priorities, and Needs, Hirtshals, Denmark, 2–5 June.

Brown, L.R., Lenssen, N., and Kane, H. (1995): *Vital Signs 1995: The Trends That Are Shaping Our Future* (New York, W.W. Norton).

Delgado, C.L., and Courbois, C. (1997): 'Changing Fish Trade and Demand Patterns in Developing Countries and Their Significance for Policy Research', Markets and Structural Studies Division Paper 18 (Washington, D.C., International Food Policy Research Institute).

FAO (Food and Agriculture Organization of the United Nations) (1996a): *Food, Agriculture, and Food Security: Developments Since the World Food Conference and Prospects*, World Food Summit Technical Background Document 1 (Rome, FAO).

FAO (Food and Agriculture Organization of the United Nations) (1996b): *Food Security and Nutrition*, World Food Summit Technical Background Document 5 (Rome, FAO).

FAO (Food and Agriculture Organization of the United Nations) (1996c): *Investment in Agriculture: Evolution and Prospects*, World Food Summit Technical Background Document 10 (Rome, FAO).

FAO (Food and Agriculture Organization of the United Nations) (1996d): *Mapping Undernutrition – An Ongoing Process*, Poster prepared for the World Food Summit.

FAO (Food and Agriculture Organization of the United Nations) (1996e): *Rome Declaration on World Food Security and World Food Summit Plan of Action* (Rome, FAO).

FAO (Food and Agriculture Organization of the United Nations) (1997a): FAOSTAT database. <http://faostat.fao.org/default.htm>. Accessed May.

— (1997b): FAOSTAT database, <http://faostat.fao.org>, accessed August and September 1997.

Hazell, P. (1995): 'Technology's Contribution to Feeding the World in 2020', in *A 2020 Vision for Food, Agriculture, and the Environment: Speeches Made at an International Conference* (Washington, D.C., International Food Policy Research Institute).

IFPRI (International Food Policy Research Institute) (1995): *A 2020 Vision for Food, Agriculture, and the Environment: The Vision, Challenge, and Recommended Action* (Washington, D.C., IFPRI).

Pinstrup-Andersen, P. (1994): *World Food Trends and Future Food Security*, Food Policy Report (Washington, D.C., International Food Policy Research Institute).

Pinstrup-Andersen, P., R. Pandya-Lorch, M.W. Rosegrant (1997): *The World Food Situation: Recent Developments, Emerging Issues, and Long-Term Prospects*, 2020 Vision Food Policy Report (Washington, D.C., International Food Policy Research Institute).

Rosegrant, M.W., Agcaoili-Sombilla, M., and Perez, N.D. (1995): *Global Food Projections to 2020: Implications for Investment*, Food, Agriculture, and the Environment Discussion Paper 5 (Washington, D.C., International Food Policy Research Institute).

Rosegrant, M.W., Sombilla, M.A., Gerpacio, R.V., and Ringler, C. (1997): 'Global Food Markets and U.S. Exports in the Twenty-First Century', paper presented at the Illinois World Food

and Sustainable Agriculture Program Conference on 'Meeting the Demand for Food in the Twenty-first Century: Challenges and Opportunities for Illinois Agriculture', Urbana-Champaign, 27 May. International Food Policy Research Institute, Washington, D.C.

UN (United Nations) (1995): *World Urbanization Prospects: The 1994 Revisions* (New York, UN).

UN (United Nations) (1996): *World Population Prospects: The 1996 Revisions* (New York, UN).

UN ACC/SCN (United Nations Administrative Committee on Coordination/Sub-committee on Nutrition) (1992): *Second Report on the World Nutrition Situations*, vol. 1 (Suffolk, England, The Lavenham Press Ltd. for the United Nations ACC/SCN Secretariat).

von Braun, J., Bouis, H., Kumar, S., and Pandya-Lorch, R. (1992): *Improving Food Security of the Poor: Concept, Policy, and Programs* (Washington, D.C., International Food Policy Research Institute).

World Bank (1997): *World Development Indicators* (Washington, D.C., World Bank).

World Bank (1997): *World Development Report 1997* (New York, Oxford University Press for the World Bank).

REFLECTIONS ON THE GLOBALIZATION OF MARKETS: THREATS AND OPPORTUNITIES

ALBERTO QUADRIO CURZIO

1. Foreword

The study-week on "science for survival and sustainable development" places at its centre, in our opinion, the statement of the organisers that "the world is facing major threats from the expansion of human activities ... and [from the] destabilization of economies and social order, ... with each passing year undermining our ability to maintain a sustainable and productive world into the 21st Century and beyond. Human society is increasingly recognizing such threats" and therefore defensive measures are taken but "the destabilizing factors prevail, and the scale of possible catastrophes is rapidly growing" (Pontificia Accademia delle Scienze, 1999). Furthermore, the quoted statement says that in order to analyse and forecast such phenomena and in order to prepare scientific initiatives to place them under the control of ethical principles and public policies, these problems will be addressed by different "experts" on mathematical and theoretical modelling, on specific critical phenomena, and on ethical principles. I consider myself an economist who is expert on specific critical phenomena, and I will expound these reflections utilising an approach more closed to public policy actions. In other words an institutional-applied economic approach, instead of an economic modelling one. Of course models have a major prominence and role. But sometimes reality shows in a simple way what is going on.

The globalization of the economies is an extensive and contemporary phenomenon (for instance see Arcelli, 1997; Onofri, 1997) and I will sum up the following subjects, considered from an economist's point of view:

a. technological innovation and diffusion;
b. multinational firms and investments;
c. international trade and world GDP.

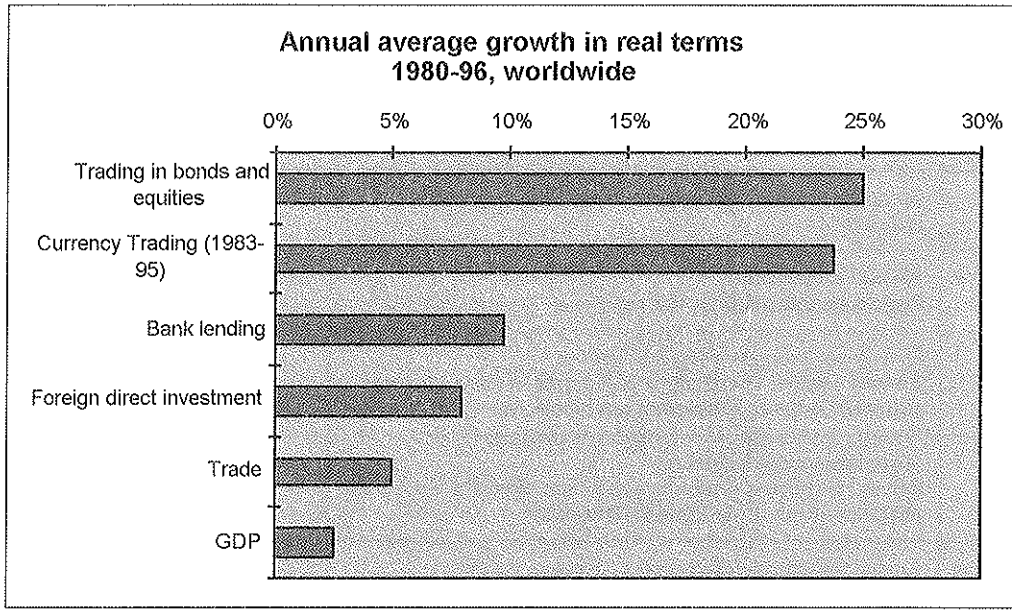
Economic phenomena also have political consequences because the dynamics of globalization are reducing the sovereignty of national States (Berger and Dore, 1998; Strange, 1998; Wade, 1998). Therefore it is increasingly urgent that we have supranational institutions which are capable of combining freedom of actors with the efficacy of rules. In the following sections I will comment on the tables and shortly explain some concepts which I discussed in detail in an essay of mine (Quadrio Curzio, 1999), to which I refer the reader for more precise bibliographical references.

2. What is Economic Globalization?

Economic globalization is broadly defined as the growing interdependence and integration of different economic systems on a world scale through the growing transactions of goods, services, financial capital, and technological diffusion (IMF, 1997). The world economy seems to be changing from a situation where the national economies were dominant and maintained their relationships at the level of international trade to a situation where the markets are global and where the national boundaries are practically irrelevant because firms loose their national roots and operate on worldwide scale. A transition from national economies, connected through international relationships, to a global economy, is underway, even if the world economy is certainly not yet fully globalized at present. As the transition from national economies to an international and global economy is not yet complete, it would be better to talk an of inter-globalized economy. The degree of globalization can be illustrated easily by two sets of data.

The first set refers to the dynamics. Figure 1 describes the dynamics of globalization on the basis of the annual average growth in real terms of four types of magnitudes during the period 1980-1996 (*The Economist*, 18 October 1997). The first set, composed of three data, represents transnational financial magnitudes which grow at an annual maximum percentage of 25%. The second set shows foreign direct investments which grow at an annual percentage near to 8%. The third one refers to international trade which grows at an annual percentage rate near to 5%. The last indicates GDP which grows at an annual percentage rate of 2.5%. The conclusion is that transnational financial and real investment activities are growing much faster than international trade and world GDP.

The second set of data has a static and structural nature and illustrates the degree of globalization of productive sectors as percentages of world GDP (Fraser and Oppenheim, 1997).



Source: *The Economist*, 18-10-1997, p. 104, based on BIS and IMF data; our graphic approximation.

Fig. 1. The Dynamics of Inter-Globalization (our title).

From table 1 it seems that 23% of world GDP is produced by global firms in global markets, 15% by firms and markets in accelerating globalization, 50% by mainly local and national firms and markets, while 12% is generated by national public services (government services). If these set of data are realistic we can say that globalization is proceeding rapidly, and this development has many consequences: the allocation of economic resources on a world scale is more effective because savings go where the returns are higher; the instability of worldwide financial capital movements can determine crises; the relationships between multinational firms, the local State, and markets can be positive for growth but also can be difficult and conflicting, although never dangerous because these firms promote economic and technological development; the possibility of a single national State to "govern" economic globalization is remote; the new sopranational institutions can create a climate of peaceful cooperation unknown to the national States. Therefore, there are positive and negative phenomena. In order to understand this process we will consider some aspects related to technology and investments.

Table 1. *Degrees of Inter-Globalization* [a].

	Size of industry based on world GDP 1995, US$ trillion [b]			
	Category Industries	GDP	%	
1	*Physical commodities* Petroleum, mineral ores, timber	2,0	23%	*Globalized*
2	*Scale-driven business goods and services* Aircraft engines, construction equipment, semiconductors, airframes, shipping, refineries, machine tools	1,0		
3	*Manufactured commodities* Refined petroleum products, aluminium, specialty steel, bulk pharmaceuticals, pulp, specialty chemicals	2,8		
4	*Labour skill-/productivity-driven consumer goods* Consumer electronics, personal computers, cameras, automobiles, televisions	0,9	15%	*Accelerating globalization*
5	*"Brandable" largely deregulated consumer goods* Soft drinks, shoes, luxury goods, pharmaceuticals, movie production	0,5[c]		
6	*Professional business services* Investment banking, legal services, accounting services, consulting services	2,5		
7	*"Hard to brand" globally, largely regulated consumer goods and services* Food, personal financial services, television production, retail distribution channels	6,3	50%	*Early globalization still local*
8	*Local (unbranded goods and services* Construction materials, real estate, funeral homes, education, household services, medical care, utilities	6,4		
9	*Government services* Civil servants, national defense	3,0	12%[d]	*Structural local*[d]
	Total	25,3		

[a] our title; [b] our caveat on trillion definition: 1 US$ trillion corresponds to 1000 US$ billions; [c] provisional; [d] percentage data and definition not present in the original source.

Source: G. Fraser e J. Oppenheim, *McKinsey Quarterly*, 1997, n. 2, p. 176. Based on: *World Development Report* (World Bank), McGraw-Hill/DRI *World Economic Outlook* 1996, United Nations *1995 National Income Accounts*, McKinsey analysis.

3. Economic Aspects of Technological Innovation and Diffusion

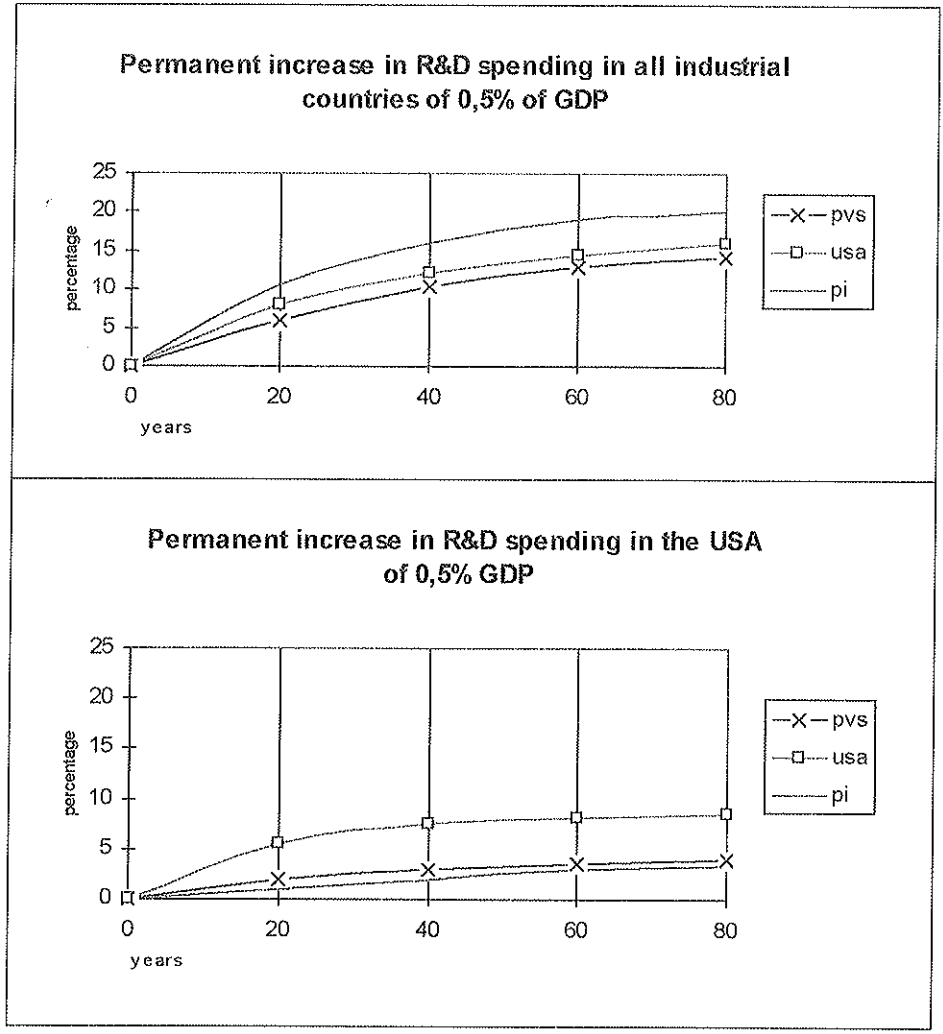
Technology seems to be the driving force of globalization through research and development; through activity of firms and international trade; and through financial and banking activity. Let us consider two aspects.

Firstly, in the opinion of many experts, the world economy is undergoing a new revolution which seems to be deeper and wider than the industrial revolution: the information and communication technologies revolution (for instance see Bensasson S., 1997; World Bank, 1998). This new revolution changes the organisation of firms and markets, often makes national boundaries irrelevant, spreads information world-wide instantaneously, and generates network integrated systems.

There are many reasons to agree with this view and all are rather simple. Let us consider for instance the capital market of some Internet firms. New firms, like Microsoft, have a market capitalization greater than General Electric, America on Line has a market capitalization greater than Boeing, with a price-earnings ratio higher than IBM. Of course we know that many experts consider these phenomena as speculative bubbles, but we also know that it is difficult to define Nasdaq's firms as the products of simple speculation. In fact, the core of this revolution is the fact that the cost of computer processing power has been falling dramatically in real terms over the past couple of decades, while computing processing power has increased (*The Economist*, 18 October 1997; IMF, 1997). And the diffusion of information technology, like Internet networks, is advancing rapidly all over the world.

In our opinion, this information technology revolution is parallel to the global spillovers of R&D, as shown in a very simple way in figure 2. According to an IMF study (IMF, 1997), a permanent increase in R&D spending of 0,5% of GDP in the USA has effects both on the potential output of the USA and on other countries trading with the USA. The same is true for an increase on the R&D spending of other industrialized countries than the USA. These increases in R&D spending have positive effects on the potential GDP of other countries through international trade, and therefore less developed countries can also benefit.

The conclusion is that information and communication technologies, the diffusion of other technology, and R&D have highly positive effects on the growth of potential GDP in developing countries aswell. From this point of view, national States must increase R&D and liberalize commerce, which are a vehicle for diffusion of technology and growth.



Source: IMF, *World Economic Outlook*, 1997, p. 49, our graphic approximation.

Fig. 2. Impact of R&D investments on GDP Direct and Indirect Effects (our title).

4. Multinational Firms and Investments

Let us now consider globalization in terms of the real and productive economic profile of multinational and global firms and foreign direct investments. Let us consider the quality and quantity of these phenomena, defining generically multinational firms as those with "remarkable" assets, employment, and production outside their home country.

Briefly, many reasons have brought about the worldwide spread of firms and they may be summarised here (Reich, 1993; *The Economist*, 22 November 1997; Lafay, 1998): productive factors (in order to have lower input costs for labour and safer inputs for raw materials); goods (in order to avoid customs); markets (in order to supply better consumers); organisations (vertical integration, economies of scale, merger and acquisitions); systems (in order to follow client firms and to keep pace with competitors). Multinational firms now have not only large but also medium-small size dimensions, which can utilise information and communication technologies.

Multinational firms in 1995 sold US$7 trillion thanks to their "foreign" firms; in 1996 their total stock of foreign direct investment was more than US$3 trillion (*The Economist*, 22 November 1997). For example, let us consider the fifteen largest worldwide companies by foreign assets in 1995 shown in table 2.

Six of these companies belong to the automobile industry (Ford, General Motors, Volkswagen, Toyota, Nissan, Daimler-Benz); four to the energy industry (Royal Dutch-Shell, Exxon, Elf Aquitaine, Mobil); two to the electronics and electrical (equipment) industry (General Electric, Abb); one to the chemicals industry (Bayer), one to the food industry (Nestlé); and one to the computers industry (Ibm). Seven of these companies are European, six are American, and two are Japanese. This table shows foreign assets, foreign sales and foreign employment as a percentage of these total entities. *The Economist* maintains that there is not real globalization because the average multinational produces more than two-thirds of its output and locates two-thirds of its employees in its home country (*The Economist*, 22 November 1997).

In our opinion what is relevant is that among these companies nine have more than 50% of their assets abroad, five more than 70%. Ten have more than 60% of their sales abroad and four more than 70%. Lastly, seven have more than 50% of their employment abroad and three more than 70%.

Of course, as we have already said, measuring globalization is not easy but many other elements show that is increasing. Let us consider the dynamics of foreign direct investments (FDI) and the comparison between

Table 2. *Top 15 TNCs by foreign assets, 1995.*

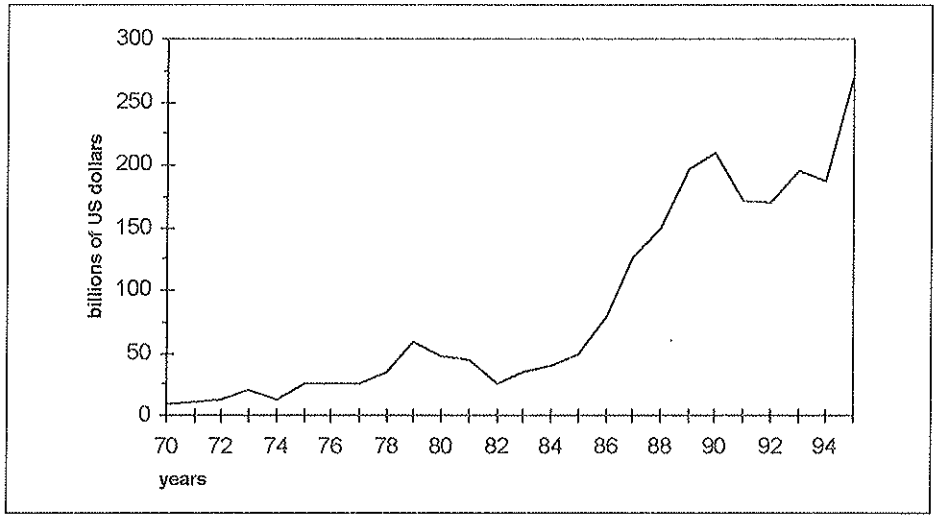
Company	Industry	Foreign assets as % of total	Foreign sales as % of total	Foreign employment as % of total
Royal Dutch/Shell	Energy	67,8	73,3	77,9
Ford	Automobile	29,0	30,6	29,8
General Electric	Electronics	30,4	24,4	32,4
Exxon	Energy	73,1	79,6	53,7
General Motors	Automobile	24,9	29,2	33,9
Volkswagen	Automobile	84,8	60,8	44,4
IBM	Computers	51,9	62,7	50,1
Toyota	Automobile	30,5	45,1	23,0
Nestlé	Food	86,9	98,2	97,0
Bayer	Chemicals	89,8	63,3	54,6
ABB	Electrical equipment	84,7	87,2	93,9
Nissan	Automobile	42,7	44,2	43,5
Elf Aquitaine	Energy	54,5	65,4	47,5
Mobil	Energy	61,8	65,9	52,2
Daimler-Benz	Automobile	39,2	63,2	22,2

Source: *The Economist*, 22.11.1997, p. 108, based on UNCTAD data.

FDI outflows and domestic investment (DI), which reveal different paths, as shown in figures 3 and 4.

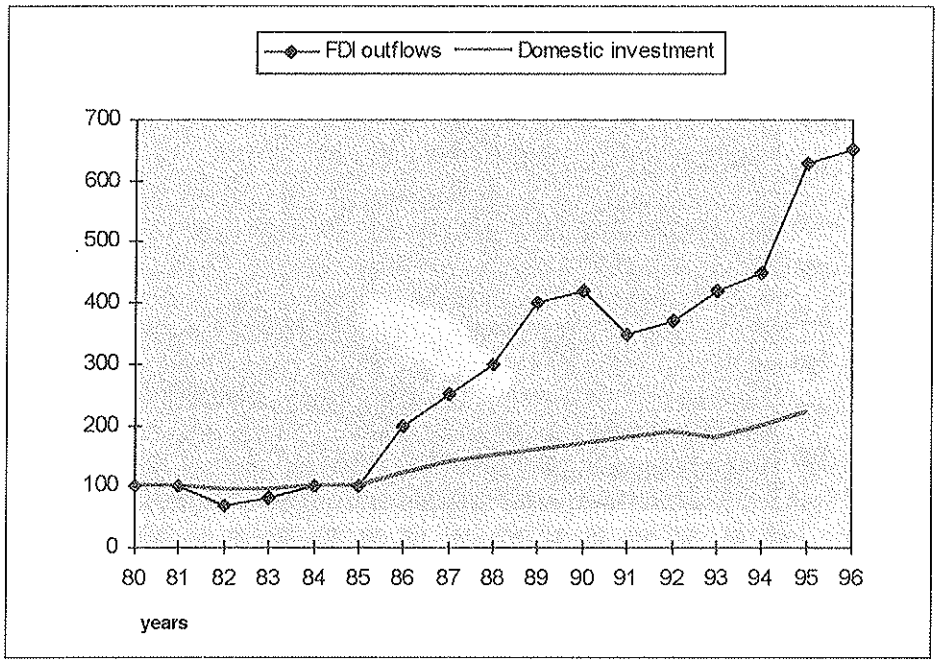
Considering the FDI outflows of eleven industrialized countries it is clear that from the '70s to the mid-'90s there was a highly positive and increasing trend. This is even clearer when we compare FDI outflows and DI. Equalising both at 100 in 1980 (*The Economist*, 22 November 1997), the growth of FDI is stronger than that of DI, even if their total size is around 6,5% of domestic investment.

To conclude the debate about multinational or global firms is complex and has sometimes also been harsh because some think that they exploit local situations in developing countries, while others think they contribute to development and to technology diffusion. In our opinion the complexity of the problem remains even if this second judgement is truer, and truer also considering that many multinational corporations are moving towards a network system based on good agreements with local authorities and firms.



* Countries included are Belgium, Canada, France, Germany, Italy, Japan, the Netherlands, Sweden, Switzerland, the United Kingdom, and the United States. These countries account for over 80 percent of the total outward stock of foreign direct investment. Data prior to 1971 exclude Germany; 1975, France and Belgium; 1977, Japan, and 1983, Switzerland.
Source: IMF, *World Economic Outlook*, 1997, p. 61; our graphic approximation.

Fig. 3. Foreign Direct Investment of the Main Industrialized Countries (1970-1995) (our title).*



Source: *The Economist*, 22.11.1997, p. 108, based on UNCTAD data; our graphic approximation.

Fig. 4. World Foreign Direct Investment and Domestic Investement (our title).

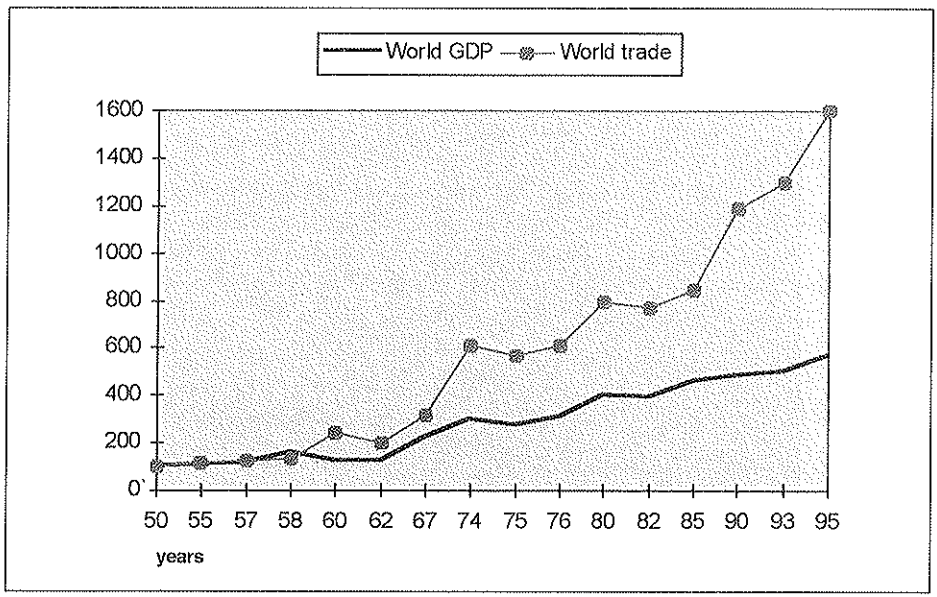
5. International Trade and World GDP

We must not forget that international trade remains a very important feature in globalization and that a remarkable contribution to it is given by the intrafirm trade of multinational corporations (WTO, 1996).

The increase of international trade is also a remarkable consequence of the liberalization achieved by rules established by Gatt (now WTO).

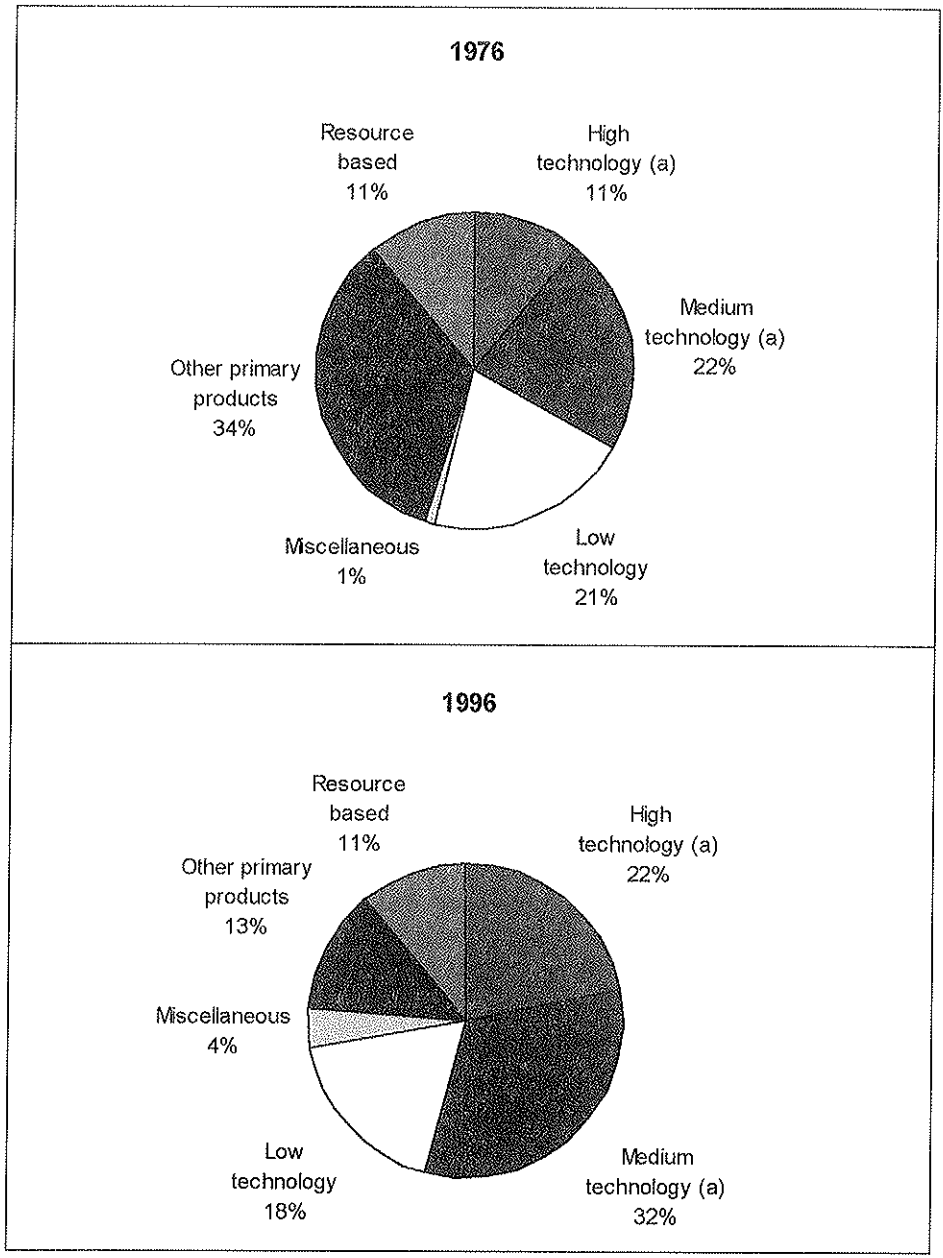
Two considerations seem useful. The first consideration concerns the dynamics of world trade and world GDP, which are clearly shown in figure 5. If we put merchandise world trade and world GDP at 100 in 1950, by 1995 the former grows to 1600, and the latter to 600 (*The Economist*, 8 November 1997). It is true that the ratio of world exports to world GDP is only 15%, but we must remember it has been growing since 1950, when it was only 7%, and that some countries (for example Germany, the United Kingdom, France, and Italy) have a ratio near 30%.

Against the thesis of rapid globalization is the situation of the prices of internationally traded goods. An obvious consideration is that homogeneous prices for homogeneous goods are the most important qualitative indicators of the integration of a market. But there is still large deviation



Source: *The Economist*, 8.11.1997, p. 99, based on WTO data; our graphic approximation.

Fig. 5. World trade and World GDP (1950=100) (our title).



Source: World Bank, *World Development Report*, 1998, p. 28, based on World Bank, COMTRADE database.

Fig. 6. Goods in International Trade by Level of Technological Intensity (medium and high-technology goods are those requiring intensive R&D as measured by R&D expenditure).

from a tendency to homogeneous prices for many internationally traded goods. For this reason, we can say that the world markets of goods are not as integrated as the national ones (IMF, 1997).

The second consideration concerns the structure of international trade, and in particular the international trade composition by level of technological intensity.

According to figure 6, in 1976 the share of trade in medium and high-technology goods was 33% of total international trade, by 1996 this share had grown to 54%, and in particular high-technology goods doubled their share: from 11% in 1976 to 22% in 1996. This involves two important developments: world merchandise trade is changing greatly and the role of firms, which produce and trade these international goods is increasing more and more. So it is reasonable to think that the role of global firms with intensive R&D in their productions is expanding.

6. Some Conclusions: the Necessity for Sopranational Institutions and Rules

In the process of economic globalization there are some advantages and some disadvantages. In our opinion the potential advantages in terms of technological diffusion, the international division of labour, saving and production allocation, are greater than the disadvantages in terms of financial instability, exploitation of local situations in developing countries, and unemployment in developed countries. But the risks of financial instability are remarkable even if the crises of the past ten years have been overcome (for instance see Fazio, 1997).

It is therefore necessary to promote a growing role for sopranational, international, co-operative agreements because a single national State cannot establish rules for the global economy and compel actors to respect them. The national State must also have sound internal fiscal-financial conditions and look after the soundness of the banking sector.

Firms, banks, and other economic agents need clear and respected rules for world markets. Competition is not anarchy. But freedom with rules is a powerful engine of development, and protectionism demonstrated that it was a disaster, and not only in extreme cases, during the great and tragic dictatorships of this century when some States cancelled the freedom of human beings and economic initiative. In the new sopranational and international system, particular attention must be paid to less developed and developing countries in terms of financial conditions (and in relation to external debt), the exportation of their commodities (avoiding protection-

ism in developed countries), and the transfer to them of technology. History seems to teach that the combination of markets and institutions, of freedom, subsidiarity and solidarity, are the only successful ways to achieve long-term development both in terms of economic and social welfare and in terms of the promotion of civilisation and human dignity.

BIBLIOGRAPHY

Arcelli, M. (ed.) (1997): *Globalizzazione dei Mercati e Orizzonti del Capitalismo* (Roma-Bari, Laterza).

Bensasson, S. (1997): 'Lo Sviluppo Tecnologico e la Società Informatica: l'Impatto della Tecnologia sulla Produzione e sul Mercato', in P. Onofri (ed.), *Lo Scenario Mondiale e il Futuro dell'Economia Italiana*, (Bologna, Il Mulino), pp. 39-47.

Berger, S., and Dore, R. (eds.) (1998): *Differenze Nazionali e Capitalismo Globale* (Bologna, Il Mulino).

The Economist, 18 October 1997, *Schools Brief, One World?*, pp. 103-104.

The Economist, 8 November 1997, *Schools Brief, Trade Winds*, pp. 99-100.

The Economist, 22 November 1997, *Schools Brief, Worldbeater, Inc.*, pp. 108-109.

Fazio, A. (1997): 'Sviluppo Economico e Mercato Globale', in M. Arcelli (ed.), *Globalizzazione dei Mercati e Orizzonti del Capitalismo* (Roma-Bari, Laterza), pp. 3-13.

Fraser, J., and Oppenheim, J. (1997): 'What's New about Globalization?', *The McKinsey Quarterly*, (2), pp. 168-179.

IMF (1997): *World Economic Outlook, Globalization, Opportunities and Challenges,* May 1997 (Washington D.C.).

Lafay, G. (1998): *Capire la Globalizzazione* (Bologna, Il Mulino).

Onofri, P. (ed.) (1997): *Lo Scenario Mondiale e il Futuro dell'Economia Italiana* (Bologna, Il Mulino).

Pontificia Accademia delle Scienze, 12 January 1999 (mimeo).

Quadrio Curzio, A. (1999): 'Globalizzazione: Profili Economici', *Rendimenti Morali, Accademia dei Lincei*, s. 9, v. 10, pp. 297-321.

Reich, R. (1993): *L'Economia delle Nazioni* (Milano, *Il Sole-24 Ore*, Libri).

Strange, S. (1998): *Chi Governa l'Economia Mondiale? Crisi dello Stato e Dispersione del Potere* (Bologna, Il Mulino).

Wade, R. (1998): 'La Globalizzazione e i suoi Limiti', in S. Berger and R. Dore (eds.), *Differenze Nazionali e Capitalismo Globale* (Bologna, Il Mulino), pp. 77-112.

World Bank (1997): *The State in a Changing World* (New York, Oxford University Press).

World Bank (1998): *World Development Report. Knowledge for Development* (New York, Oxford University Press).

WTO (1996): *Trade and Foreign Direct Investment*, GATT Focus, December 1996.

ECOLOGICAL SYSTEMS AND ECONOMIC INSTITUTIONS

PARTHA S. DASGUPTA

Given the theme of our study-week, I can take the seriousness of the problem of environmental security (and the many other things which go with it) for granted. So I will focus elsewhere, on certain technical matters concerning ecosystem dynamics and their implications for our understanding of the efficacy of different economic institutions. If you were to find it surprising that there could be connections between what might appear to be widely different fields of inquiry, I would be pleased, for this would mean that I have something of interest to talk about.

1. Local vs. Global Constraints

Ecologists' findings suggest that a near-fifty percent increase in world population, allied to a doubling of gross world product per head, by the year 2040 or so, would create substantial additional "stresses" in both local and global ecosystems.[1] For example, global "demand" for food could easily double over the period 1990-2030, with two-and-a-half to three-fold increases in the poorest countries. Of particular concern are Asia and Africa where, over the next fifty years, plant-derived food-energy requirements are expected to increase by a factor of 2.3 and 5, respectively, with a more-than-sevenfold increase expected in some countries (Pinstrup-Andersen, 1994; Crosson and Anderson, 1995; FAO, 1996). And these figures do not include the inevitable increases in the demand for non-food commodities that would accompany increases in GNP. This is why much attention has been given in recent years to global environmental problems.

---

[1] See, for example, Vitousek *et al.* (1986) and the Symposium on the scale of human activity in *Science*, 20 July 1997.

The prospects for a suitable response to the predicted increases in the scale of the human enterprise depend on our ability to manage constraints on the supplies of production inputs and on the environmental consequences of the use of these inputs. These constraints are not present uniformly across the globe. Moreover, local problems of production and distribution can be difficult to counter even when global supplies are adequate, because the purchasing power of certain groups of people (viz. the poor) is weak. To ask merely whether global production of goods and services can be increased to meet future demands in a sustainable way misses much of the question. For example, food scarcity manifests itself locally, so efforts to alleviate it must be tailored to local circumstances. To do otherwise is akin to doctoring a sick person on the basis of global health statistics.

Correct diagnosis of the problems that lie at the population-consumption-environment nexus is usually a *local* matter, even though appropriate treatment may require regional and global support. For example, soil erosion may not currently be a serious threat to global agricultural capacity, but at local levels in various parts of the world it presents major problems to the people affected. Similarly, decisions concerning fertility, education, child-care, food, work, health-care, and the use of the local natural-resource base are in large measure reached and implemented within households, who face constraints that are shaped in part by national and international policy. The influence of household decisions are felt through local interactions (e.g. intra-village and village-town trades), and thence "upward" globally. Recent work has identified a variety of circumstances that are shaped by positive feedback mechanisms, driving poverty, hunger, fertility, resource degradation, and civic disconnection at the local level, even while national (and not merely global) income is rising.[2] This means in particular that there can be bifurcations among otherwise similarly located people, so that even while some groups enjoy greater and greater economic comforts, others at best stagnate or, worse still, spiral downward. This suggests that if we are to obtain reliable projections of global economic prospects, we need to adopt local, contemporary perspectives. It also reminds us that environmental problems constitute not only the ones that are aired during international negotiations – there are a myriad of local environmental problems in constant need of attention by local people.

---

[2] See Dasgupta and Mäler (1991), Dasgupta (1993, 1995, 1998) and Cleaver and Schreiber (1994).

2. Non-linear Processes, Substitution Possibilities and the Market Mechanism

A major achievement of modern economics has been to show the efficacy of the competitive market mechanism for the allocation of resources. Central to the market mechanism is the role played by prices. Under ideal conditions market prices would reflect resource scarcities in an accurate way. Prices would get bid up if traders observed that resources were becoming more scarce and would get bid down if large quantities of substitutes were discovered. So market prices are public (and aggregated) signals of what are typically a myriad of private pieces of information, known only locally. Resources are allocated efficiently when prices reflect social scarcities accurately.

However, modern economics has shown that the market mechanism could be expected to work well only if the transformation of goods and services into further goods and services is governed by linear processes. Prices are able to reflect resource scarcities in an accurate manner under such circumstances. But when one speaks of "stress on ecosystems" and "positive feedback mechanisms", as I did in the previous section, one refers to systems characterized by non-convex processes (more generally, to non-linear processes) and so to the possible bifurcations they are subject to. We have already noted that such processes can govern both global and local systems. Even if a given global ecosystem were not to show signs of stress, local ones could, and often do, display such signs. There are also extant records of local ecosystems having collapsed in the past. Prices cannot be expected to reflect resource scarcities well when ecological transformations are governed by non-linear processes. This explains the title of my essay.

The assumption of linearity in economic transformation possibilities is related to the idea that for every commodity that can be transacted, there are close substitutes lying waiting. The latter assumption, if true, would imply that even as constraints increasingly make their presence felt on any one resource base, humanity could move to other resource bases. The enormous additions to the sources of industrial energy (successively human power, animal and wind power, timber, coal, oil and natural gas and, most recently, nuclear) that have been realized are a prime historical illustration of this possibility.

The assumption of linearity continues to be reasonable in many spheres of activity, but it becomes sorely stretched when applied to those that encroach ecosystems (be they local or global) on a large scale. The services provided by an ecosystem are dependent on the composition of biota and the abiotic processes at work. Here it is important to distinguish between the resource base that comprises an ecosystem (its structure) and the serv-

ices the ecosystem provides (its functions).[3] Degradation of the resource base (e.g. destruction of populations) not only affects the volume and quality of those services; it also challenges an ecosystem's "resilience", which is the capacity of the system to absorb disturbances without undergoing fundamental changes in its functional characteristics.[4] If a system loses its resilience, it can flip to a wholly new state when subjected to even a small perturbation (see, e.g. Wilson, 1992; Holling *et al.*, 1995; Walker, 1995; Levin *et al.*, 1998). One way to interpret an ecosystem's loss of resilience is to view it as having moved to a new stability domain. Sudden changes in the character of shallow lakes (e.g. from clear to eutrophied water), owing to increases in the input of nutrients, provide one class of examples (Scheffer, 1997; Carpenter, Ludwig and Brock, 1998); the transformation of grasslands into shrublands, consequent upon non-adaptive cattle-management practices, provides another (Perrings and Walker, 1995).[5]

Closely related is the concept of "biodiversity". Even today it is a popular belief among the general public that the utilitarian value of biodiversity is to be located solely in the potential uses of genetic material (e.g. for pharmaceutical purposes). Preservation of biodiversity is seen as a way of holding a diverse portfolio of assets with uncertain payoffs. But as I understand it, biodiversity *appropriately conceived* is associated positively with a system's resilience and with its "productivity" (as measured by the flow of energy and the internal cycling of nutrients). This implies that ecosystems must harbour biodiversity if they are to *be* productive (Tilman, 1997). There is then the important corollary that, to invoke the idea of substitutability among natural resources in commodity production in order to play down the utilitarian importance of biodiversity, as people frequently do (e.g. Simon, 1981, 1994), is a wrong intellectual move. Biodiversity would appear to be a necessary condition for substitute ecosystem services to *be* available. Its importance cannot be downplayed by the mere hope that there are substitutes lying in wait.[6]

---

[3] Daily (1997) contains a useful collection of essays on the character of these services.

[4] So, in concentrating on functional (as opposed to structural) characteristics, I am taking an entirely utilitarian view of ecosystems.

[5] Recovery can be costly, in some cases impossible. In short, such flips can in many cases be regarded as irreversible. The mathematics of "relaxation phenomena" offers a formal account of what the intuitive notion of irreversibility amounts to. On this, see for example, Levin (1999).

[6] Recall the famous analogy in Ehrlich and Ehrlich (1981) relating species in an ecosystem to rivets in an airplane: one by one, perhaps, species may disappear and not be missed. Eventually, however, the cumulative effect of loss of biodiversity will lead to the crash of ecosystem functioning, just as the cumulative loss of redundant rivets will lead to the crash of an airplane.

3. Market Failure Owing to its Non-Existence

So there is a link between ecological non-linearities and the ability of the market mechanism to function well. But there is another broad class of reasons why we should not expect markets to function well in regard to the environmental resource-base, be they global or local.

The reason is that for many environmental resources markets simply do not exist. In some cases they do not exist because the costs of negotiation and monitoring are too high. One class of examples is provided by economic activities that are affected by ecological interactions involving long geographical distances (e.g. the effects of deforestation in the uplands on downstream activities hundreds of miles away); another, by large temporal distances (e.g. the effect of carbon emission on climate in the distant future, in a world where forward markets are non-existent because future generations are not present today to negotiate with us). Then there are cases (e.g. the atmosphere, aquifers, and the open seas) where the nature of the physical situation (viz. the migratory nature of the resource) makes private property rights impractical and so keeps markets from existing; while in others (e.g. biodiversity; see Perrings *et al.*, 1994), ill-specified or unprotected property rights prevent their existence, or make markets function wrongly even when they do exist. In short, environmental problems are often caused by market failure.

Problems arising from an absence of forward markets for "transactions" between the now and the distant future are no doubt ameliorated by the fact that we care about our children's well-being and know that they in turn will care for theirs, and so on, down the generations. This means, by recursion, that even if we do not care directly about the well-being of our distant descendents, we do care about them indirectly. However, there is a distinct possibility that our implicit concern for the distant future via such recursion is inadequate. This is why many economists have argued that market rates of interest do not reflect social discount rates.[7] In short, market failure involves not only misallocation of resources in the present, but also misallocation across time.

Since markets cannot be relied upon to provide us with prices that would signal true environmental scarcities, there is a need to develop techniques for enabling us to do so. A great deal of work in environmental and resource economics has been directed at discovering methods for estimat-

[7] Lind (1982), Arrow *et al.* (1996) and Dasgupta, Maler, and Barrett (1998) contain accounts of these considerations.

ing notional prices, often called "accounting prices" by economists, that could be used by decision-makers.[8] But for the most part practical methods have been developed for estimating the accounting prices of amenities – relatively few for the multitude of ecosystem services that constitute our life-support system.[9] This is why the indicators of social well-being frequently in use (e.g. gross national product per head (GNP), life expectancy at birth, and the infant survival rate) do not reflect deteriorations in the environmental resource base which might be expected to be associated with economic activities.

4. Indicators of Social Well-Being: GNP vs. NNP

To illustrate, consider that such indices of the standard of living as GNP per head pertain to commodity production, they do not fully take into account the use of natural capital in the production process. So statistics on past movements of gross product tell us nothing about the environmental-resource base. They do not say if, for example, increases in GNP per head are not being realized by means of a depletion of this base (for example, if increases in agricultural production are not being achieved by "mining" the soil, lowering of water tables and impairment of other ecosystem services). Such impairment can easily go unrecorded because, as I observed earlier, the use of ecosystem services all too often involves transactions that are not mediated by an effective "price system". So, for example, if, when drawing water from an aquifer, individual farmers were to ignore the effect of their extraction on others' *future* extraction costs owing to a lowering of the water table, the social cost of agricultural production would exceed the farmers' private costs. Even though each farmer would, typically, impose only a small additional cost on others, the sum of the costs imposed by each on all others could well be substantial. This means that it is possible for the real costs of agricultural production to exceed the market prices of agricultural produce. Indeed, it is even possible for their market prices to decline over time even while the real cost of production is rising (see accompanying figure, where a hypothetical case is shown). By concentrating on current-welfare measures, such as GNP, market prices of agricul-

[8] Resources for the Future, a research institution in Washington, D.C., has been a pioneer in this field of inquiry.

[9] An example is the "contingent valuation method" (see Mitchell and Carson, 1989), which was devised for eliciting in quantitive terms, through the use of questionaires, the extent to which people value environmental amenities.

tural produce, and life expectancy at birth, economists, journalists and political leaders have, for the most part, wrongly bypassed the links that exist between population growth, increased material output, and the state of the natural-resource base.

Over the years environmental and resource economists have demonstrated how national product, if it is to function effectively as an index of the standard of living, should include the value of changes in the environmental resource-base that occur each year.[10] They have also shown that, when it is properly defined, this index, which measures *net* national product (NNP), takes into account the effect of changes in stocks of natural capital on future consumption possibilities. But note that it is possible for an economy to record increases in GNP per head over an extended period even while NNP per head is declining (see the accompanying diagram). We should be in a position to say if this has not been happening in the various regions of the world. But the practice of national-income accounting has lagged so far behind its theory, that we have little idea of what the facts have been. It is therefore entirely possible that time trends in GNP per head give us a singularly misleading picture of movements of the true standard of living.

To put the matter otherwise, current-day estimates of GNP are biased because the accounting value of changes in the stocks of natural capital are not taken into account. What is usually called NNP merely deducts the depreciation of manufactured capital (e.g. buildings and machinery) from GNP. The NNP whose use we are advocating here goes farther by also deducting the depreciation of natural capital. In other words, NNP estimates used in policy debates are biased because a biased set of prices is in use. As their accounting prices are not available, environmental natural resources on site are frequently imputed to have no value. This amounts to regarding the depreciation of environmental capital as of no consequence. But as these resources are scarce goods, their accounting prices are positive. So, if they depreciate, there *is* a social loss. It means that profits attributed to projects that degrade the environment are greater than the social profits they generate. Estimates of their rates of return are higher than their true rates of return. Wrong sets of investment projects therefore get selected, in both the private and public sectors: resource-intensive projects look better than they actually are. It should be no surprise then that installed technologies are often unfriendly towards the environment. This is probably so

[10] See, for example, Mäler (1974), Dasgupta and Heal (1979), and Dasgupta and Mäler (1998a).

especially in poor countries, where environmental legislations are usually neither strong nor effectively enforced.

The extent of such bias in investment activities will obviously vary from case to case, and from country to country. But it can be substantial. In their work on the depreciation of natural resources in Costa Rica, Solorzano *et al.* (1991) have estimated that in 1989 the depreciation of three resources – forests, soil, and fisheries – amounted to about 10 percent of gross domestic product and over a third of gross capital accumulation.

One can go further: the bias extends to the prior stage of research and development. When environmental natural resources are underpriced (in the extreme, when they are not priced at all), there is little incentive on anyone's part to develop technologies that would economise on their use. So the direction of technological research and technological change are systematically directed against the environment. Often enough in consequence, environmental "cures" are sought once it is perceived that past choices have been damaging to the environment, whereas "prevention" would have been the better choice.[11]

5. Enviromnental Taxes and Subsidies

Taxes, subsidies, even quantity restrictions, are but the other side of accounting prices. We should interpret taxes on the use of environmental services to be the difference between market and accounting prices. One way of improving the performance of markets. is to impose regulations on resource users; for instance, quantitative restrictions on the discharge of toxic substances and on activities that can curtail ecosystem services. As noted above, the services that underpin their production and perform other life support functions of cleansing, recycling, renewal, and protection from lethal hazards, are not priced in the market place. Therefore society, collectively, needs to control such things as the use of insecticides in ways that damage natural pest control or pollination services, use of rivers or oceans for sewage disposal, or the release of chemicals that assault the ozone shield. Strictly enforced quotas can also be imposed on fish harvests.

---

[11] To give an example, Chichilnisky and Heal (1998) have compared the costs of restoring the integrity of the Catskills Watershed, in New York State, to the costs of replacing the natural water purification services the ecosystem has provided in the past by building an $8 billion water-purification plant. They have shown the overwhelming economic advantages of preservation over cure: independent of the other services the Catskills watershed provides, and ignoring the annual running costs of $300 million for a filtration plant, the capital costs alone showed a more than 6-fold advantage for investing in the natural-capital base.

Another way to improve the performance of markets is to introduce systems of taxes on "harmful" activities and subsidies for "beneficial" ones. They would include pollution charges, taxes on fish harvested, and on activities that would degrade ecosystem services (for example, release of carbon dioxide or deforestation that reduces flood control). The idea here is to choose the rates of taxation to reflect the harmful spillovers these activities inflict on those not engaged in them. By the same token, subsidies can be (and are) given for the acquisition and dissemination of knowledge. Environmental taxes and subsidies, if chosen judiciously, would bring the prices of environmental goods and services in line with their value to society, and thereby improve the performance of the market system.

Each of the two schemes has advantages and disadvantages over the other. Regulations may produce faster, surer results. But environmental taxes, when properly designed, often can reduce pollution more effectively than regulations. In addition, there is a presumption that tax revenues would enable the government to reduce other, economically distorting taxes (e.g. taxes on earned income). There is thus a possibility that the imposition of environmental (or "green") taxes could yield a "double dividend": less pollution and a more efficient economy. This rhetorical phrase has been much used in recent years to persuade governments to impose such taxes. But in the real world, the double dividends may not materialize. The pollution tax may drive up prices, cost jobs, reduce the tax base, and require increases in such distorting taxes as those on income. Such cascading consequences can plague those trying to manage both economic and ecological systems.

6. Valuing Changes to the Environment vs. Valuing the Environment

It is worth emphasising that the purpose of estimating environmental accounting prices is not to value the *entire* environment; rather, it is to evaluate the benefits and costs associated with *changes* made to the environment due to human activities. Prices have significance only when there are potential exchanges from which choices have to be made (for example, when one has to choose among alternative investment projects). Thus, the statement that a particular act of investment can be expected to degrade the environment by, say, 1 million dollars annually, has meaning, because it says among other things that if the investment were not to be undertaken, then, other things being the same, humanity would enjoy an additional 1 million dollars of benefits in the form of environmental services. The statement also has operational significance: the estimate could (and should!) be used for calculating the rate of return attributable to the investment in question.

However, there is no meaning in such a string of words as that world-wide the flow of environmental services is currently worth 33 trillion US dollars annually (Costanza *et al.*, 1997). The reason it has no meaning is that if environmental services were to cease, life would not exist. Therefore, to say that world-wide the flow of environmental services is currently worth some 33 trillion US dollars annually has no more meaning than the assertion that humanity is worth 33 trillion US dollars annually. But what would it mean to say that humanity is annually worth this figure, or for that matter *any* figure? Who would be there to receive those dollar benefits if humanity were to exchange its existence for them? It is not so much the crudeness of the estimate as its meaninglessness that is the fatal flaw.

The point is important. It does not do to defend such estimates against criticism, as one of the authors has done, by insisting that they are a mere "first cut" at what is a very difficult measurement problem (Robert Costanza, as reported in Masood and Garwin, 1998, p. 430). The point of the criticism is not that the authors got their estimates "wrong", that further work would get the estimate more nearly right; the point of the criticism is that the very idea of arriving at a figure for the dollar worth of global environmental services is meaningless.[12]

Nor does it do to say that publicizing the estimate has done a service by bringing environmental concerns onto the agenda of public discourse. Wrong arguments, no matter how well intentioned the cause, can only undermine progress and confound isues. The essential point is that life would be extinguished if all environmental services were to cease. One does not have to engage in economic accounting to arrive at this simple truth.

7. Institutional Failure and Ecosystem Destruction

To sum up: markets cannot be relied upon to generate correct signals of resource scarcity not only because the nature of the situation surrounding ecosystem services can keep certain crucial markets from existing, but also because ecological processes that are involved in the transformation of goods and services into other goods and services frequently involve "non-linear" processes. Non-linearities make their presence particularly felt when the systems in question are under stress (e.g. when they approach "thresholds"). This being so, we should not expect markets to generate those sig-

[12] Massood and Garwin (1998, p. 430) appear to think that it is meaningless only in the context of what they call "neo-classical economics". In the text we have stressed the fact is that it is meaningless in any context.

nals which would alert us to impending shifts in the stability regimes of ecosystems. Human populations have on occasions been unable to prevent suffering from unexpected flips in their local ecosystems because of this.

This said, ecosystem degradation can occur not only because of market failure, it can occur also because of bad government policies (e.g. because of wrong tax policies).[13] We may put the matter more generally: an underlying cause of environmental degradation is *institutional failure*. If I have stressed market failure in this essay, it is because we economists understand the market mechanism better than non-economists, and because we understand it better than we understand most other resource allocation mechanisms. But the various types of institutional failure I have alluded to pull in different directions and are together not unrelated to an intellectual tension between the concerns people share about such matters as mean global warming and acid rains, which sweep across regions, nations and continents; and about those matters (such as, for example, the decline in firewood or water sources) that are specific to the needs and concerns of the poor in as small a group as a village community. Environmental problems present themselves differently to different people. Some people identify environmental problems with population growth, while others identify them with wrong sorts of economic growth. Then there are others who view them through the spectacle of poverty. Each of these visions is correct. There is no single environmental problem; rather, there is a large collection of them, some global, many local.

Over the past many years now, environmental and resource economists have responded to this fact by identifying desirable institutional reforms in a case-by-case manner. Alterations to prevailing structures of property rights, the imposition of environmental and resource taxes, regulations, local-community control, and various other devices that change individual and group incentives have been much discussed. Contrary to what is frequently suggested in popular writings on environmental matters, the tools of modern economics are *not* restricted to the study of convex systems. Many of the lessons drawn have been put into use, most especially in western industrial countries.

However, far less work has been done on the economics of local ecosystems in poor societies. There is a reason for this. Because economic systems often do not generate signals that would alert the public of growing resource scarcity, it can be a very difficult matter for those who suffer from

[13] Binswanger (1991) has argued that government policies in Brazil regarding agricultural income and land ownership have in the past provided incentives for deforestation in the Amazon basin.

the economic consequences of the scarcity to get an environmental problem placed on the agenda of public discourse. In poor countries, for example, there are strong links between household poverty, local environmental deterioration, and a weak political voice (see e.g. Dasgupta, 1997). As in many other aspects of life, the political economy of the matter, and in particular *governance*, is at the heart of many environmental problems.

REFERENCES

Arrow, K.J., Cline, W.R., Mäler, K.-G., Munasinghe, M., Squitieri, R., and Stiglitz, J.E. (1996): 'Intertemporal Equity, Dscounting, and Economic Efficienty', in IPCC, *Climate Change 1995: Economic and Social Dimensions of Climate Change*, Contribution of Working Group III to the Second Assessment Report of the Intergovernmental Panel on Climate Change, edited by J.P. Bruce, H. Lee and E.F. Haites (Cambridge, Cambridge University Press).

Binswanger, H. (1991): 'Brazilian Policies that Encourage Deforestation in the Amazon', *World Development*, 19 (7), pp. 821-829.

Carpenter, S.R., Ludwig, D., and Brock, W.A. (1998): 'Management of Eutrophication for Lakes Subject to Potentially Irreversible Change', Discussion Paper, Beijer International Institute of Ecological Economics, Stockholm.

Chichilnisky, G., and Heal, G. (1998): 'Economic Returns from the Biosphere', *Nature*, 391, pp. 629-630.

Cleaver, K.M., and Schreiber, G.A. (1994): *Reversing the Spiral: the Population, Agriculture, and Environment Nexus in Sub-Saharan Africa* (Washington, D.C., World Bank).

Costanza, R., d'Arge, R., de Groot, R., Farber, S., Grasso, M., Hannon, B., Limburg, K., Naeem, S., O'Neill, R.V., Paruelo, J., Raskin, R.G., Sutton, P., and van den Belt, M. (1997): 'The Value of the World's Ecosystem Services and Natural Capital', *Nature*, 387, pp. 253-260.

Crosson, P., and Anderson, J.R. (1995): 'Demand and Supply: Trends in Global Agricolture', *Food Policy*, 19, pp. 105-110.

Daily, G. (ed.) (1997): *Nature's Services: Societal Dependence on Natural Ecosystems* (Washington, D.C., Island Press).

Daily, G., Dasgupta, P., Bolin, B., Crosson, P., du Guerny, J., Ehrlich, P., Folke, C., Jansson, A.-M., Jansson, B.-O., Kautsky, N., Kinzig, A., Levin, S., Mäler, K.-G., Pinstrup-Andersen, P., Siniscalco, D., and Walker, B. (1998): 'Food Production, Population Growth and Environmental Security', Discussion Paper, Beijer international Institute of Eclogical Economics, Stockholm. Forthcoming, *Science*, 1998.

Dasgupta, P. (1993): *An Inquiry into Well-Being and Destitution* (Oxford, Clarendon Press).

Dasgupta, P. (1995): 'The Population Problem: Theory and Evidence', *Journal of Economic Literature*, 33 (4), pp. 1879-1902.

Dasgupta, P. (1998): 'The Economics of Poverty in Poor Countries', *Scandinavian Journal of Economics*, 100 (1), pp. 41-68.

Dasgupta, P., and Heal, G. (1979): *Economic Theory and Exhaustible Resources* (Cambridge, Cambridge University Press).

Dasgupta, P., Levin S., and Lubchenco, J. (1998): 'Economic Pathways to Ecological Sustainability: Challenges for the New Millenium', mimeo., Department of Evolutionary Biology, Princeton University.

Dasgupta, P., and Mäler, K.-G. (1991): 'The Environment and Emerging Development Issues', *Proceedings of the Annual World Bank Conference on Development Economics, 1990* (Supplement to the *World Bank Economic Review* and the *World Bank Research Observer*), pp. 101-132.

Dasgupta, P., and Mäler, K.-G. (1998a): 'Analysis, Facts and Prediction', *Environment and Development Economics*, 3, pp. 504-510.

Dasgupta, P., and Mäler, K.-G. (1998b): 'Decentralization Schemes, Cost-Benefit Analysis, and Net National Product as a Measure of Social Well-Being', forthcoming, *Environment and Development Economics*.

Ehrlich, P.R., and Ehrlich, A.H. (1981): *Extinction: the Causes and Consequences of the Disappearance of Species* (New York, NY, Random House).

FAO (1996): *World Food Summit Technical Document 4* (Rome, Food and Agriculture Organization).

Holling, C.S. (1986): 'The Resilience of Terrestrial Ecosystem: Local Surprise and Global Change', in W.C. Clark and R.E. Munn (eds.), *Sustainable Development of the Biosphere* (Cambridge, Cambridge University Press).

Holling, C.S., Schindler, D.W., Walker, B.W., and Roughgarden, J. (1995): 'Biodiversity in the Functioning of Ecosystems: An Ecological Synthesis', in C. Perrings *et al.*, *Biodiversity Loss: Economic and Ecological Issues* (Cambridge, Cambridge University Press).

Levin, S.A. (1999): *The Fragile Dominion: Complexity and the Commons* (Reading, MA, Addison Wesley Longman).

Levin, S.A., Barrett, S., Aniyar, S., Baumol, W., Bliss, C., Bolin, B., Dasgupta, P., Ehrlich, P., Folke, C., Gren, I.M., Holling, C.S., Jansson, A., Jansson, B.-O., Martin, D., Maler, K.-G., Perrings, C., and Sheshinsky, E. (1998): 'Resilience in Natural and Socioeconomic Systems', *Environment and Developmental Economics*, 3, pp. 225-236.

Lind, R.C. (ed.) (1982): *Discounting for Time and Risk in Energy Planning* (Baltimore, Johns Hopkins University Press).

Lubchenco, J., Olson, A.M., Brubaker, L.B., Carpenter, S.R., Holland, M.M., Hubbell, S., Levin, S.A., MacMahon, J.A., Matson, P.A., Melillo, J.M., Mooney, H.A., Peterson, C.H., Pulliam, R., Real, L.A., Regal, P.J., and Risser, P.G. (1991): 'The Sustainable Biosphere Initiative: An Ecological Research Agenda', *Ecology*, 72, pp. 371-412.

Mäler, K.-G. (1974): *Environmental Economics: A Theoretical Enquiry* (Baltimore, Maryland, Johns Hopkins University Press).

Massod, E., and Grawin, L. (1998): 'Costing the Earth: When Ecology Meets Economics', *Nature*, 395, pp. 426-430.

Mitchell, R.C., and Carson, R.T. (1989): *Using Surveys to Value Public Goods: the Contingent Valuation Method* (Washington, D.C., Resources for the Future).

PCAST (1998): 'Teaming with Life: Investing in Science to Understand and Use America's Limiting Capital', PCAST Panel on Biodiversity and Ecosystems (Washington, D.C.).

Perrings, C. *et al.* (1995): *Biodiversity Loss: Economic and Ecological Issues* (Cambridge, Cambridge University Press).

Perrings, C., and Walker, B.W. (1995): 'Biodiversity Loss and the Economics of Discontinuous Change in Semi-Arid Rangelands', in C. Perrings *et al.*, *Biodiversity Loss: Economic and Ecological Issues* (Cambridge, Cambridge University Press).

Pinstrup-Andersen, P. (1994): 'World Food Trends and Future Food Security', Food Policy Report, International Food Policy Research Institute, Washington, D.C.

Scheffer, M. (1997): *The Ecology of Shallow Lakes* (New York, Chapman Hall).

Simon, J.L. (1981): *The Ultimate Resource* (Princeton, Princeton University Press).

Simon, J.L. (1994): 'Debate Statement', in N. Myers and J. Simon, *Scarcity or Abundance? A Debate on the Environment* (New York, W.W. Norton, 1994).

Solorzano, R. *et al.* (1991): *Accounts Overdue: Natural Resource Depreciation in Costa Rica* (Washington, D.C., World Resources Institute).

Tilman, D. (1997): 'Biodiversity and Ecosystem Functioning', in G. Daily (ed.), *Nature's Services: Societal Dependence on Natural Ecosystems* (Washington, D.C., Island Press).

Vitousek, P.M., Ehrlich, A., Ehrlich, P., and Matson, P. (1986): 'Human Appropriation of the Products of Photosynthesis', *BioScience*, 36 (6), pp. 368-372.

Walker, B.H. (1995): 'Rangeland Ecology: Managing Change in Biodiversity', in C. Perrings *et al.*, *Biodiversity Conservation* (Dordrecht, Kluwer).

Wilson, E.O. (1992): *The Diversity of Life* (Cambridge, MA, Harvard University Press).

World Bank (1992): *World Development Report* (New York, Oxford University Press).

ENERGY PRUDENCE

WALLACE S. BROECKER

INTRODUCTION

A debate rages regarding the significance of the changes in climate which will occur during the next century as the result of the ongoing buildup of greenhouse gases in the Earth's atmosphere. A large majority of scientists involved in atmosphere and ocean research support the conclusions of the IPCC report which state that if unabated, this buildup will result in a significant warming of the planet with consequent changes in rainfall, storminess, and soil moisture. These changes, which will intensify over the course of the next century, may adversely impact the production of food and will certainly pose an additional threat to the already stressed wildlife on our planet.

One prominent atmospheric scientist, MIT's Richard Lindzen, strongly opposes this view. He correctly points out that a significant warming will occur only if the primary forcing (by CO_2, CH_4, N_2O, CFCs) is amplified by an increase in the atmosphere's water vapor content. In the absence of this water vapor feedback, a tripling of the atmosphere's CO_2 content would lead to only a 1.8°C global warming. The general circulation models (GCMs) employed by atmospheric scientists have in common that the water vapor content of the atmosphere rises in proportion to the vapor pressure of water (i.e., about 7 percent per °C) and thereby through its large infra red absorption capacity generates an amplification of primary warming by a factor of about 2.5, thus vaulting the warming for a CO_2 tripling to about 4.5°C. Lindzen (1982) is convinced that existing general circulation models (GCMs) models do not distribute water vapor correctly and, in particular, that the water vapor content of the air descending over the desert regions of the world will decrease rather than increase. As these regions constitute the atmosphere's major radiator, this decrease would tend to null the primary CO_2 warming.

My view lies at the opposite pole from that of Lindzen. I believe that the results obtained using GCMs are the best guide we have to the future. Further, in addition to the gradual warming predicted by these GCMs, I fear that our planet's climate system may ultimately undergo an abrupt reorganization. If the Earth's response to the buildup of greenhouse gases is as large as the GCMs predict and if the buildup were to triple the pre-industrial CO_2 content (i.e. 3x280 or 840 ppm), there is a distinct possibility that warming and increased precipitation in the polar regions will lead to a disruption of the ocean's large scale thermohaline circulation. By analogy to events recorded in ice cores, in mountain moraines and in rapidly accumulating marine sediments, such a disruption would bring about a large and abrupt change in the climate of our planet.

Expected Magnitude of the Greenhouse Buildup

Before discussing the evidence in support of a possible greenhouse-triggered reorganization, a few words regarding the expected magnitude of the anthropogenic buildup of the atmosphere's greenhouse capacity are in order. While it must be kept in mind that currently the impact of excess methane, nitrous oxide and CFCs roughly match that of CO_2, it is CO_2 emissions that pose the major future hazard for they will prove to be the most difficult to rein in. We now emit 6.8 gigatons of carbon (GtC) as CO_2 per annum. To this must be added about one GtC per year resulting from deforestation. With an expected increase in population to 9 or 10 billion and with the expected increase in the standard of living of people in developing countries, were fossil fuels to remain our primary source of energy, then it is likely that the CO_2 emissions will rise well above 10 GtC per annum before 2050 A.D. As shown in figure 1, were emissions to average 10 GtC over the entire course of the 21st century, then in the absence of a significant enhancement of storage in terrestrial biomass (i.e., trees and soil humus) the CO_2 content would reach well into the danger zone. By this, I mean that when the greenhouse contributions of CH_4, N_2O and CFCs are taken into account the greenhouse capacity of the atmosphere will surpass the level at which models suggest that thermohaline circulation would be seriously impacted (see Manabe and Stouffer, 1993 and Stocker and Schmittner, 1997). Even if by the end of the next century enhanced storage of carbon in the terrestrial biosphere reached the generous magnitude of 200 GtC, the greenhouse content of the atmosphere would still reach into the danger zone. In order to avoid entry into this zone, we would have either turn to non-fossil fuel sources to supply roughly half of the energy

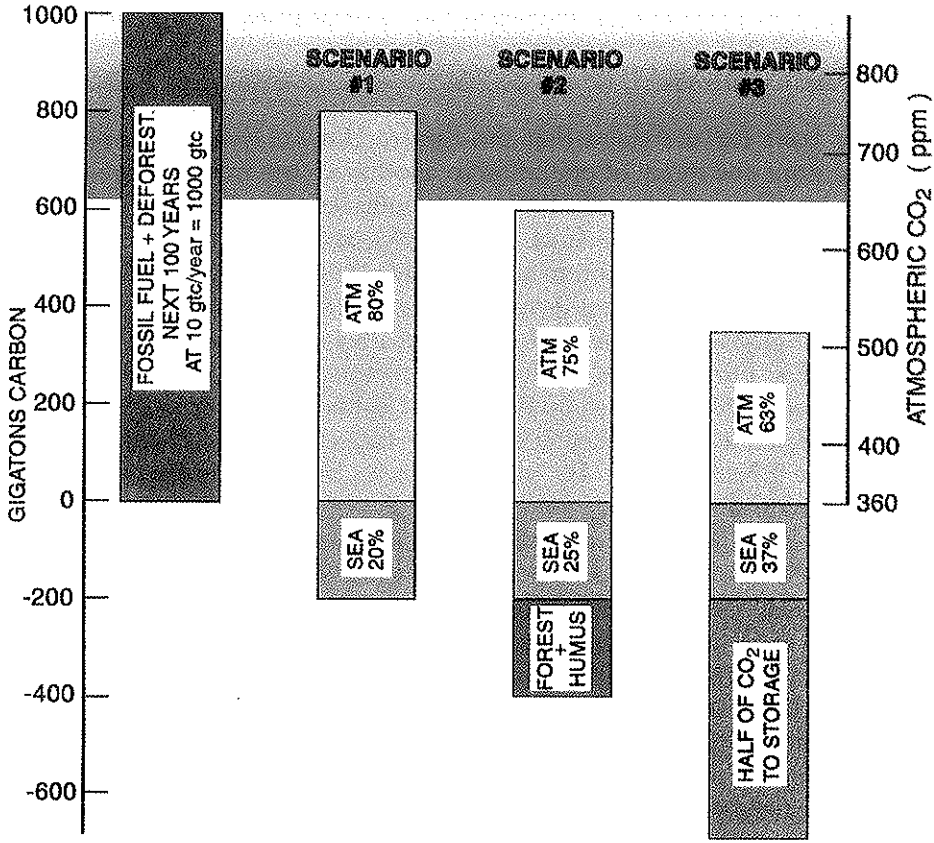


Fig. 1. If over the next century 1000 GtC are released into the atmosphere as the result of fossil fuel burning and deforestation, then in the absence of significant greening of the terrestrial biosphere roughly 80 percent will remain in the atmosphere raising its CO_2 content to about 750 ppm. If we are lucky and greening driven by fixed nitrogen and carbon dioxide increases storage in the terrestrial biosphere by as much as 200 GtC, CO_2 would increase to about 640 ppm. The only way by which the CO_2 content could be held below 500 ppm would be to capture and store more than half of the CO_2 generated or to substitute non-fossil-fuel energy systems for at least half of the conventional sources. When the contribution of CH_4, N_2O and the CFCs are taken into account, a CO_2 buildup to more than 600 parts per million puts us in the danger zone with regard to a shutdown of thermohaline circulation.

needed in the twenty-first century or we would have to sequester roughly half of the CO_2 generated by fossil fuel burning over the course of the twenty-first century.

The Climatic Record

The basis for the claim that the Earth's climate system is capable of jumping from one mode of operation to another comes from records stored in polar ice (Dansgaard *et al.*, 1993), in rapidly deposited marine sediments (Behl and Kennett, 1996; Schulz *et al.*, 1998), and in the moraines formed by mountain glaciers (Denton and Hendy, 1994). Taken together, these records provide convincing evidence that these switches were abrupt, strong, and global. By 'abrupt', I mean that they were completed in two to four decades. Furthermore, during the transition interval climate flickered (Taylor *et al.*, 1993) much as do fluorescent lights when they are turned on. By 'strong', I mean that they eclipsed, by far, any climate change experienced during historic time. By 'global', I mean that their impacts were felt everywhere on the planet.

As I have published several papers summarizing this evidence (see Broecker, 1997b), I will present here only a brief recap. During the course of the last glacial period, the Earth experienced 20 or so millennial-duration oscillations in climate between a very cold state and an intermediate cold state. The last of these major chills has been dubbed the Younger Dryas (Y. D.). Its abrupt ending came 11,500 years ago. Since then, except for one-century duration cooling centered at 8,200 years ago (Alley *et al.*, 1997), the Earth's climate system has remained locked in its interglacial mode. The birth of agriculture occurred early in the present interglacial (and probably in response to the reduction in middle eastern rainfall brought about by the demise of glacial climates). Hence, civilization as we know it developed during a period of unusual climate quiescence. No mode change has marred its progress.

Due to a combination of results of annual layer-counting in Greenland's ice cores, radiocarbon dating of moraines and sediments, and thorium-uranium dating of corals and speliothems, we can speak with confidence about the chronology of the events of the last 130,000 years These results provide a firm absolute chronology covering the entire duration of the last major glacial-interglacial cycle. In particular, the radiocarbon method allows us to correlate events during the last 40,000 years across the entire planet.

Much of the information comes from 2 three-kilometer-long ice cores located close to the geographic center of the Greenland ice cap. Drilled at

sites separated by 30 kilometers, these cores provide records agreeing to the finest detail back to 110,000 years ago (i.e., back to roughly the middle of the last interglacial). The oxygen isotope record in the ice itself provides the pattern of the temperature changes (see figure 2). The structure of the temperature profile measured in the bore hole itself provides a means of calibrating the isotope record. It shows that the mean air temperature over Greenland's ice plateau was on the average about 16°C colder than now during glacial time (Cuffey *et al.*, 1994). The dust content in the ice varied in concert with the isotopes (Mayewski *et al.*, 1994). It ranged up to 50 times higher than today's concentration. As isotope fingerprinting (via radiogenic daughter nuclides of the elements lead, strontium and neodymium) demonstrates that this dust originated in Asia's Gobi Desert (Biscaye *et al.*, 1997), this concordance requires that the storminess over Asia underwent jumps in frequency and intensity in exact concert with Greenland's air temperature changes. Furthermore, the input of dust clearly flickered during the transitions (Taylor *et al.*, 1993). Finally, the methane content of air trapped in bubbles in the ice shows sharp changes (Chappellaz *et al.*, 1993; Brook *et al.*, 1996) which have been shown to be synchronous with the dust and temperature changes (Severinghaus *et al.*, 1998). As the major source of methane during glacial time was likely to have been tropical swamps, these water bodies must have become warmer and wetter at the times of the abrupt cold to warm transitions in Greenland.

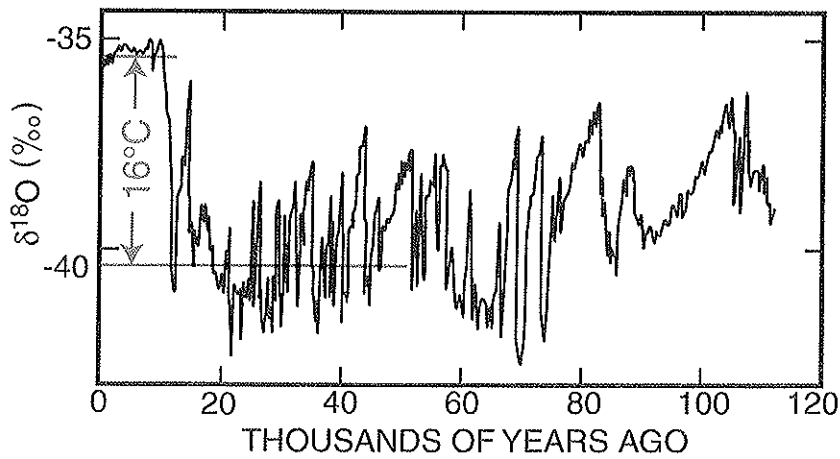


Fig. 2. The pattern of temperature change over the last 110,000 years as recorded by the ^{18}O to ^{16}O ratio in Greenland ice (Dansgaard *et al.*, 1993). The absolute temperature range has been independently determined from the temperature profile measured in the borehole (Cuffey *et al.*, 1994).

Further evidence for the widespread occurrence of Greenland's millennial duration events comes from marine sediments from areas of high sedimentation rate (i.e., >50 cm/10^3 yrs.). Three such sites, one in the Santa Barbara basin off California (Behl and Kennett, 1996), one in the Arabian Sea off India (Schulz *et al.*, 1998), and one in the Cariaco Trench off Venezuela (Hughen, personal communication), are currently bathed in oxygen-poor intermediate depth water (see figure 3). During the present interglacial, the sediments at these sites are for the most part annually layered demonstrating that oxygen is absent in the sediment pore waters. This absence prevents worms from stirring the sediment and thereby erasing the layering. At all three locations, the O_2 content of the sediment must have been considerably higher during the Younger Dryas and also at the times of millennial-duration cold extremes. This alternation in thermocline O_2 content makes clear that some combination of the rate of ventilation by O_2-rich surface waters and the rate of rain of organic matter from the overlying water at these cold

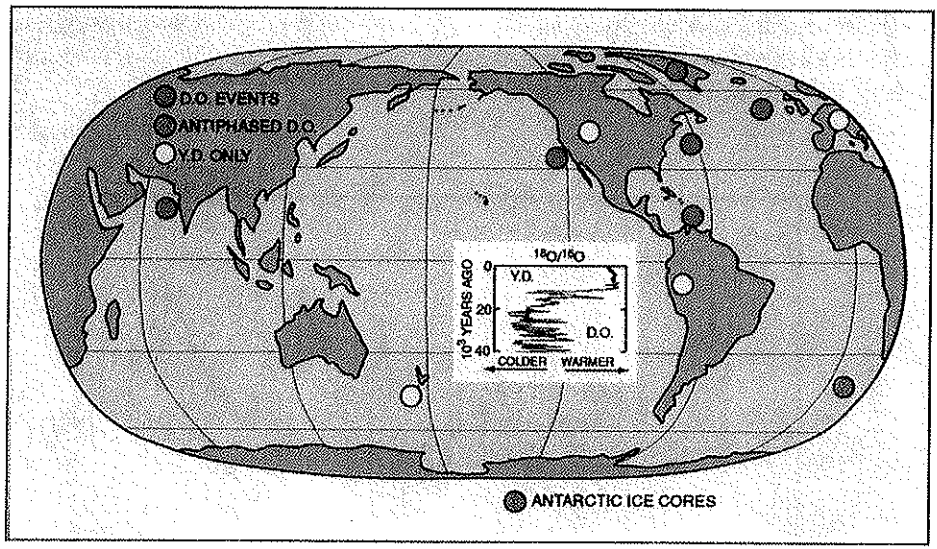


Fig. 3. The red dots depict localities on the globe where the full set of Greenland's Dansgaard Oeschger events have been identified (Behl and Kennett, 1996; Schulz *et al.*, 1998). The yellow dots depict localities where radiometric ages confirm a Younger Dryas age for mountain moraines (Gosse *et al.*, 1995; Denton and Hendy, 1994; Ivy-Ochs *et al.*, 1996) and for ^{18}O cold anomalies in Andean ice cores (Thompson *et al.*, 1998). The blue dots depict southern sites where these events are antiphased with respect to those for the rest of the globe (Blunier *et al.*, 1997 and 1998).

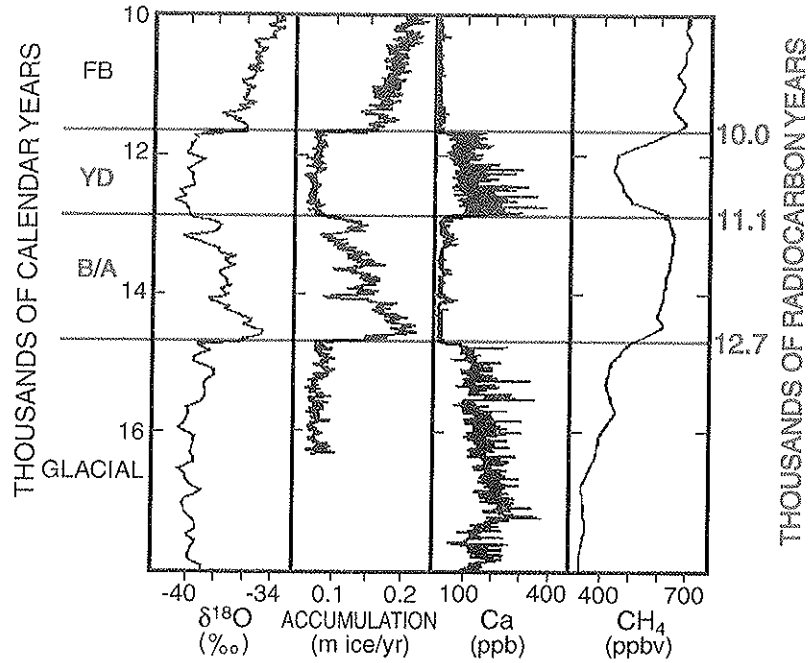


Fig. 4. Oxygen isotope (Grootes *et al.*, 1993), snow accumulation rate (Alley *et al.*, 1993), calcium component of dust (Mayewski *et al.*, 1994) and methane content of trapped air (Chappellaz *et al.*, 1993) of Summit Grenland ice cores for the deglaciation period including the Bolling-Allerod (BOA) warm and the Younger Dryas (YD) cold.

times must have to be maintained higher pore water O_2 contents. While it is not possible to say exactly how the operation of the upper oceans changed, the fact that the same temporal pattern is seen in the Pacific, Indian and Atlantic Oceans suggests that the changes were global in scale.

Detailed measurements of alkenone ratios (an excellent paleothermometer) on rapidly accumulating sediment from a site near Bermuda reveal that during the time interval 30 to 60 thousand years ago, surface water temperatures in the Sargasso Sea underwent 4 to 5°C changes in concert with the temperature swings observed in the Greenland ice core (Sachs and Lehman, in press).

Only for the period between the onset of the Bolling-Allerod warm period which brought to an end the last glacial period and the subsequent Younger Dryas cold lapse (see figure 4) do we have extensive coverage for the continents. The pollen records contained in bog and lake sediments clearly demonstrate that a profound global vegetation change occurred at

the time of the onset of the Bolling-Allerod warm. The vastness of this climate change is also recorded by closed basin lakes in desert regions. Prior to the onset of the Bolling-Allerod, Lake Victoria which straddles the equator in East Africa was bone dry (Johnson *et al.*, 1996). The lake reappeared early in the Bolling-Allerod. The nearby Red Sea had become so saline during late glacial time that planktonic foraminifera could no longer survive (Hemleben *et al.*, 1996). Then suddenly at the onset of the Bolling-Allerod warm, planktonic foraminifera reappeared. In contrast, the situation in the North America's Great Basin was exactly the opposite. During late glacial time its closed basin lakes, Bonneville and Lahontan, were as much as ten times larger in area than those of the present day remnant lakes (Benson, 1981, 1993). The shutdown of the water supply that maintained these large lakes came at the onset of the Bolling-Allerod. Hence, not only was this event marked by a pronounced global warming but also by a major redistribution of rainfall (see Broecker *et al.*, 1998).

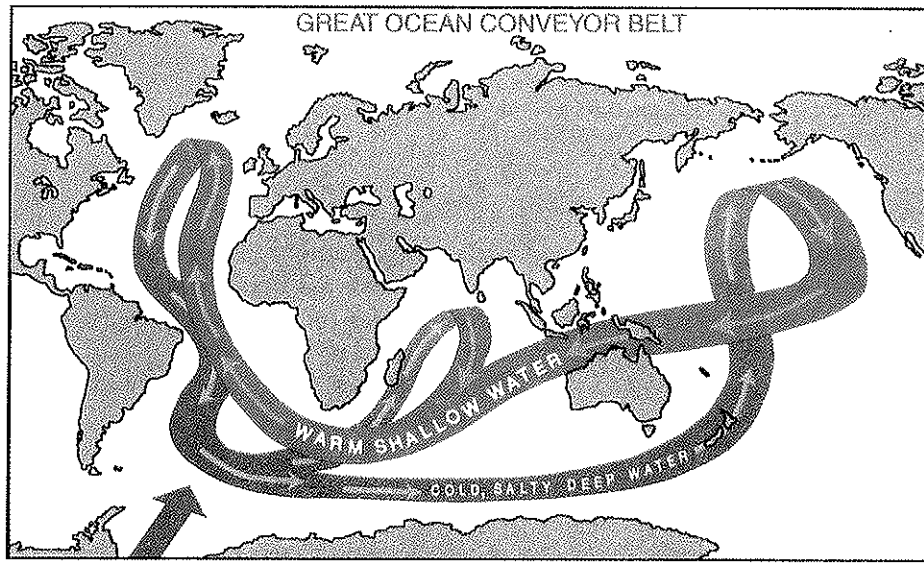


Fig. 5. A conveyor-like circulation in the Atlantic Ocean carries an enormous amount of heat to the vicinity of Iceland. Here winter cooling densifies this already salty water allowing it sink to the bottom. This newly formed deep water flows to the south passing around the southern tip of Africa where it is joined by deep waters formed in the Weddell Sea. These waters mix to form a circumpolar deep current, which swirls around the Antarctic continent. This water eventually peals off and floods the deep Indian and Pacific Oceans (Broecker, 1991).

Abrupt-Change Triggers

Only one part of the climate system, the Atlantic Ocean's conveyor circulation, has been demonstrated to have clear defined alternate modes of operation (figure 5). Model studies demonstrate that the ocean's large-scale thermohaline (thermo = cold, haline = salty) circulation can lock into more than one pattern (Marotske and Willebrand, 1991; Stocker *et al.*, 1992; Rahmstorf, 1994, 1995). The existence of these alternate patterns is possible because the deep sea can be ventilated from both high northern and southern latitudes. Which particular source locale dominates depends on the distribution of salt in surface ocean waters (the density of seawater rises with increasing salt content and with decreasing temperature). Today's ocean surface waters of the northern Pacific are so low in salt content that even when at their freezing point (i.e., –1.8°C) they do not descend more than a few hundred meters. By contrast, waters in the northern Atlantic are particularly high in salt content. When cooled to only +2°C, they become dense enough to sink to the bottom and flood southward into the circumpolar raceway surrounding the Antarctic continent. While the surface waters in the Southern Ocean are, like those in the northern Pacific, too fresh to permit them to sink to the abyss, beneath the sea ice pack covering the narrow continental margin of Antarctica brines released during the winter expansion of the ice densify the sub ice waters to the point where they spill off the shelves and sink to the abyss. Currently these waters join those emanating from the Atlantic to form a mix which flows northward into the deep Pacific and Indian Oceans.

A variety of evidence from marine sediments tells us that the currently important role of deep waters formed in the northern Atlantic was greatly lessened during glacial time. Further, that reorganizations of the ocean's thermohaline circulation were intimate parts of the abrupt change process is demonstrated by the record of ^{14}C to C ratios in the upper ocean and atmosphere (Hughen *et al.*, 1998). These ratios have been reconstructed by making radiocarbon measurements on carbon from materials of known calendar age. Such calendar ages can be obtained by counting annual layers (rings produced by trees and annual layers deposited in standing water). Radiocarbon measurements on planktonic foraminifera from varved Cariaco Trench sediments revealed that during the first 200 years of the Y. D. the $^{14}C/C$ ratio in the surface waters of the Caribbean Sea rose by an astounding 5 percent (Hughen *et al.*, 1998). Then, during the course of the remaining 1000 years of Y. D. time, the $^{14}C/C$ ratio gradually declined back to its pre Y. D. value. As the onset of the ^{14}C rise coincides exactly with the onset of the Y. D. (marked by a distinct and sharp color change in Cariaco

sediment), the only reasonable explanation for its occurrence is that the formation of new deep water in the northern Atlantic came to a halt. As today, the conveyor-like circulation in the Atlantic supplies most of the deep sea's radiocarbon and as deep sea waters house about two thirds of the radiocarbon present on our planet, this cutoff would lead to a backlogging in the atmosphere and upper ocean of radiocarbon atoms produced by the bombardment of our planet by cosmic rays. Assuming that these upper reservoirs contain about one third of the ocean-atmosphere carbon, their $^{14}C/C$ ratio would rise by 5% in 200 years. Correspondingly, the $^{14}C/C$ ratio in the deep sea would fall by 2.5% in 200 years. So, the radiocarbon record from the Cariaco Trench provides a smoking gun with regard to the intimate involvement of thermohaline circulation. The most reasonable cause for the shutdown of the Atlantic's conveyor circulation at the onset of the Y. D. is a sudden influx of fresh water into the region where deep waters are formed. Indeed, we now know that such a flood did take place. Drainage from a large lake impounded in front of the retreating Canadian ice sheet suddenly shifted from the Mississippi River to the St. Lawrence (Teller and Thorleifson, 1983). This shift was heralded by a drop in the lake's outlet permitting a large portion of the stored water to flood directly into the region where deep waters are formed.

In addition, there is evidence for a second melt water flood close to the onset of a brief Y.D.-like cold event at 8,200 years ago (Barber *et al.*, in press). This flood occurred when melt water tunneled its way through the retreating Canadian ice sheet into Hudson Bay, once again flooding the area where deep water forms. As was the case for the Y. D., this brief cold episode came to an end when the Atlantic's conveyor popped back into action.

While these two cold events which punctuated the present interglacial were very likely to have been triggered by floods of fresh water, it is difficult to believe that each of the twenty or so glacial flip flops recorded in the glacial portion of the Greenland ice record were triggered by a melt water flood. Rather, I envision that they were driven by an oscillation in the salt content of the Atlantic Ocean (Broecker *et al.*, 1990). During the periods of intense cold, precipitation over Canada and Scandinavia became locked in the growing ice sheets allowing the Atlantic to become more salty. Eventually a point was reached where the density of surface waters became high enough to permit the conveyor to snap back into action. The heat transported by the upper limb of the conveyor then changed the situation. Net growth of the ice gave way to net melting. The consequent input of fresh water drove down the salinity of Atlantic waters until the conveyor once again shutdown. This cycle repeated itself over and over again on a millennial time scale.

The Illusive Global Connection

While reorganizations of the ocean's thermohaline circulation appear to have served as the trigger for abrupt change, the link to global climate remains obscure. Studies conducted in joint ocean-atmosphere models suggest that the climate changes associated with a shutdown or rejuvenation of the Atlantic's conveyor should be confined in latitude to a belt north of Gibraltar and in longitude by the western margin of the Atlantic and roughly the Eurasian boundary. Winter temperatures in this region would be 6 to 10 degrees Celsius warmer when the conveyor is operating than when it is shut down. However, as we have seen, the impacts of the abrupt changes were global. The lowering of mountain snowlines during the Y. D. was about the same at 40°S as at 40°N (see figure 3). In order to produce such large and hemispherically symmetrical changes, it is necessary to impact the great equatorial convective systems which supply moisture to much of the world's atmosphere. But alas, to date, no one has been able to articulate even a first order conceptual model designed to link the tropics to the Atlantic's conveyor. So difficult does this task seem that some dynamists are tempted to turn the connection around and somehow trigger a change in the dynamics of the tropical atmosphere that would then propagate to the thermohaline circulation.

Prognosis for the Future

Clearly in the absence of an understanding of the mechanism which allows the Earth's climate to undergo radical changes, there is no way to predict whether by adding greenhouse gases to the atmosphere, we put the Earth at risk. However, if as I suspect, these mode switches are triggered by reorganizations of the ocean's thermohaline circulation, then one avenue we can follow is to assess at what point such a buildup might impact the production of new deep water. Clearly, as the planet warms the density of surface waters in the polar regions will decline. In addition to the thermal effect, there will be a dilution of the salt content of these waters, for on a warmer planet, there will be more precipitation. In areas poleward of 40° where precipitation exceeds evaporation, this excess will result in increased wetting. Indeed, joint ocean-atmosphere models (Manabe and Stauffer, 1993; Stocker and Schmittner, 1997) suggest that if the greenhouse capacity of our atmosphere (i.e., the joint impact of excess CO_2, CH_4 N_2O, and CFCs) were to reach the CO_2 equivalent of 750 ppm, then a substantial weakening of the Atlantic's conveyor circulation would be brought about. By analogy to past events such a weakening might lead to a jump to one of

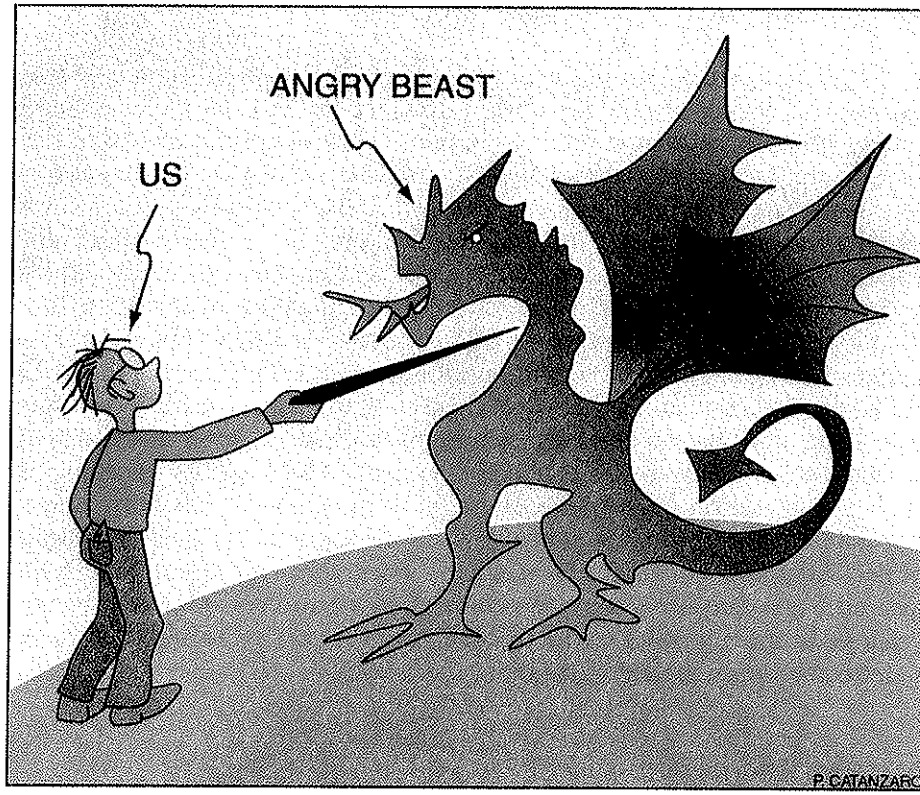


Fig. 6. The Earth's climate system has proven itself to be an angry beast. We are now poking this beast with greenhouse gases. Unfortunately, we lack the where-with-all to predict how the beast will respond.

the Earth's climate systems alternate modes. As in the past such jumps were heralded by a several decade long series of flickers, such a change would certainly at least temporarily stress our ability to feed the 9 or so billion people we expect to be present on the planet a century from now.

Conclusions

The record kept in Greenland ice and in rapidly accumulating marine sediments speaks clearly to me that the Earth's climate system is an angry beast. Through the addition of greenhouse gases to the atmosphere we are poking this beast. But as we have yet to fully understand the rules governing the beast's behavior, we cannot, as yet, predict its response to our poke.

However, prudence dictates that we take firm steps to cut back the release of CO_2 into the atmosphere. These steps will involve a combination of energy conservation, the implementation of non-fossil-fuel-based energy sources, and the capture and sequestration of CO_2 from stationary power plants.

REFERENCES

Alley, R.B., Meese, D.A., Shuman, C.A., Gow, A.J., Taylor, K.C., Grootes, P.M., White, J.W.C., Ram, M., Waddington, E.D., Mayewski, P.A., and Zielinski, G.A. (1993): 'Abrupt Increase in Greenland Snow Accumulation at the End of the Younger Dryas Event', *Nature*, v. 362, pp. 527-529.

Alley, R.B., Mayewski, P.A., Sowers, T., Stuiver, M., Taylor, K.C., and Clark, P.U. (1997): 'Holocene Climatic Instability: A Prominent, Widespread Event 8,200 Years Ago', *Geology*, v. 25, pp. 483-486.

Barber, D.C., Dyke, A., Hillaire-Marcel, C., Jennings, A.E., Andrews, J.T., Kerwin, M.W., Bilodeau, G., McNeely, R., Southon, J., Morehead, M.D., and Gagnon, J.-M., 'Forcing of the Cold Event 8,200 Years Ago by Outburst Drainage of Laurentide Lakes', in press, 1999.

Behl, R.J., and Kennett, J.P. (1996): 'Brief Interstadial Events in the Santa Barbara Basin, NE Pacific, During the Past 60 kyr', *Nature*, v. 379, pp. 243-246.

Benson, L.V. (1981): 'Paleoclimatic Significance of Lake-level Fluctuations in the Lahontan Basin', *Quaternary Research*, v. 16, pp. 390-403.

Benson, L. (1993): 'Factors Affecting ^{14}C Ages of Lacustrine Carbonates: Timing and Duration of the Last Highstand Lake in the Lahontan Basin', *Quaternary Research*, v. 39, pp. 163-174.

Biscaye, P.E., Grousset, F.E., Revel, M., Van der Gaast, S., Zielinski, G.A., Vaars, A., and Kukla, G. (1997): 'Asian Provenance of Glacial Dust (stage 2) in the Greenland Ice Sheet Project 2 Ice Core, Summit, Greenland', *Journal of Geophysical Research*, v. 102, pp. 26, 765 - 26, 781.

Blunier, T., Schwander, J., Stauffer, B., Stocker, T., Dallenbach, A., Indermühle, A., Tschumi, J., Chappellaz, J., Raynaud, D., and Barnola, J.-M. (1997): 'Timing of the Antarctic Cold Reversal and the Atmospheric CO_2 Increase with Respect to the Younger Dryas Event', *Geophysical Research Letters*, v. 24, pp. 2683-2686.

Blunier, T., Chappellaz, J., Schwander, J., Dallenbach, A., Stauffer, B., Stocker, T.F., Raynaud, D., Jouzel, J., Clausen, H.B., Hammer, C.U., and Johnsen, S.J. (1998): 'Asynchrony of Antarctic and Greenland Climate Change During the Last Glacial Period', *Nature*, v. 394, pp. 739-743.

Broecker, W.S., Bond, G., Klas, M., Bonani, G., and Wolfli, W. (1990): 'A Salt Oscillator in the Glacial North Atlantic? 1. The Concept', *Paleoceanography*, v. 5, pp. 469-477.

Broecker, W.S. (1991): 'The Great Ocean Conveyor', *Oceanography*, v. 4, pp. 79-89.

Broecker, W.S. (1997a): 'Will our ride into the Greenhouse Future be a Smooth One?', *GSA Today*, v. 5, pp. 1-7.

Broecker, W.S. (1997b): 'Thermohaline Circulation, The Achilles Heel of our Climate System: Will Manmade CO_2 Upset the Current Balance?', *Science*, v. 278, pp. 1582-1588.

Broecker, W.S. (1998): 'Paleocean Circulation During the Last Deglaciation: A Bipolar Seesaw?', *Paleoceanography*, v. 13, pp. 119-121.

Broecker, W.S., Peteet, D., and Hajdas, I., Lin, J., and Clark, E. (1998): 'Antiphasing between rainfall in Africa's Rift Valley and North America's Great Basin', *Quaternary Research*, v. 50, pp. 12-20.

Brook, E.J., Sowers, T., and Orchardo, J. (1996): 'Rapid Variations in Atmospheric Methane Concentration During the Past 110,000 Years', *Science*, v. 273, pp. 1087-1091.

Chappellaz, J., Blunier, T., Raynaud, D., Branola, J.M., Schwander, J., and Stauffer, B. (1993): 'Synchronous Changes in Atmospheric CH_4 and Greenland Climate between 40 and 8 kyr BP', *Nature*, v. 366, pp. 443-445.

Cuffey, K.M., Alley, R.B., Grootes, P.M., Bolzan, J.M., and Anandakrishnan, S. (1994): 'Calibration of the $\delta^{18}O$ Isotopic Paleothemmometer for Central Greenland, using Borehole Temperatures', *Journal of Glaciology*, v. 40, pp. 341-349.

Dansgaard, W., Johnsen, S.J., Clausen, H.B., Dahl-Jensen, D., Gundestrup, N.S., Hammer, C.U., Hvidberg, C.S., Steffensen, J.P., Sveinbjornsdottir, A.E., Jouzel, J., and Bond, G. (1993): 'Evidence for General Instability of Past Climate from a 250-kyr Ice-core Record', *Nature*, v. 364, pp. 218-220.

Denton, G.H., and Hendy, C.H. (1994): 'Younger Dryas Age Advance of Franz Josef Glacier in the Southern Alps of New Zealand', *Science*, v. 264, pp. 1434-1437.

Gosse, J.C., Klein, J., Evenson, E.B., Lawn, B., and Middleton, R. (1995): 'Beryllium-10 Dating of the Duration and Retreat of the Last Pinedale Glacial Sequence', *Science*, v. 268, pp. 1329-1333.

Grootes, P.M., Stuiver, M., White, J.W.C., Johnsen, J.J., and Jouzel, J. (1993): 'Comparison of Oxygen Isotope Records from the GISP2 and GRIP Greenland Ice Cores', *Nature*, v. 366, pp. 552-554.

Hemleben, C., Meischner, D., Zahn, R., Almogi-Labin, A., Erlenkeuser, H., and Hiller, B. (1996): 'Three Hundred Eighty Thousand Year Long Stable Isotope and Faunal Records from the Red Sea: Influence of Global Sea Level Change on Hydrography', *Paleoceanography*, v. 11, pp. 147-156.

Hughen, K.A., Overpeck, J.T., Lehman, S.J., Kashgarian, M., Southon, J., Peterson, L.C., Alley, R., and Sigman, D.M. (1998): 'Deglacial Changes in Ocean Circulation Form an Extended Radiocarbon Calibration', *Nature*, v. 391, pp. 65-68.

IPCC (1995): *Climate Change 1994, Radiative Forcing of Climate Change* (Intergovernmental Panel on Climate Change, Cambridge Press), 339 pp.

Ivy-Ochs, S., Schluchter, C., Kubik, P.W., Synal, H.-A., Beer, J., and Kerschner, H. (1996): 'The Exposure Age of an Egesen Moraine at Julier Pass, Switzerland, Measured with the Cosmogenic Radionuclides ^{10}Be, ^{26}Al and ^{36}Cl', *Eclogae Geologicae Helvetiae*, v. 89, pp. 1049-1063.

Johnson, T.C., Scholz, C.A., Talbot, M.R., Kelts, K., Ricketts, R.D., Ngobi, G., Beuning, K., Ssemmanda, I., and McGill, J.W. (1996): 'Late Pleistocene Desiccation of Lake Victoria and Rapid Evolution of Cichlid Fishes', *Science*, v. 273, pp. 1091-1093.

Lindzen, R.S., Hou, A.Y., and Farrel, B.F. (1982): 'The Role of Convective Model Choice in Calculating the Climate Impact of Doubling CO_2', *Journal of Atmospheric Science*, v. 39, pp. 1189-1205.

Manabe, S., and Stouffer, R.J. (1993): 'Century-scale Effects of Increased Atmospheric CO_2 on the Ocean-atmosphere System', *Nature*, v. 364, pp. 215-218.

Marotzke, J. and Willebrand, J. (1991): 'Multiple Equilibria of the Global Thermohaline Circulation', *Journal of Physical Oceanography*, v. 21, pp. 1372-1385.

Mayewski, P.A., Meeker, L.O., Whitlow, S., Twickler, M.S., Morrison, M.C., Bloomfield, P., Bond, G.C., Alley, R.B., Gow, A.J., Grootes, P.M., Meese, D.A., Ram, M., Taylor, K.C., and Wumkes, W. (1994): 'Changes in Atmospheric Circulation and Ocean Ice Cover over the North Atlantic During the Last 41,000 Years', *Science*, v. 263, pp. 1747-1751.

Rahmstorf, S. (1994): 'Rapid Climate Transitions in a Coupled Ocean-atmosphere Model', *Nature*, v. 372, pp. 82-85.

Rahmstorf, S. (1995): 'Bifurcations of the Atlantic Thermohaline Circulation in Response to Changes in the Hydrological Cycle', *Nature*, v. 378, pp. 145-149.

Rahmstorf, S. (1996): 'On the Freshwater Forcing and Transport of the Atlantic Thermohaline Circulation', *Climate Dynamics*, v. 12, pp. 799-811.

Sachs, J.P., and Lehman, S.J., 'Covariation of Subtropical Atlantic and Greenland Temperatures', *Science*, in press, 1999.

Schulz, H., von Rad, U., and Erlenkeuser, H. (1998): 'Correlation between Arabian Sea and Greenland Climate Oscillations of the Past 110,000 Years', *Nature*, v. 393, pp. 54-57.

Severinghaus, J.P., Sowers, T., Brook, E.J., Alley, R.B., and Bender, M.L. (1998): 'Timing of Abrupt Climate Change at the End of the Younger Dryas Interval from Thermally Fractionated Gases in Polar Ice', *Nature*, v. 391, pp. 141-146.

Stocker, T.G., Wright, D.G., and Broecker, W.S. (1992): 'The Influence of High-latitude Surface Forcing on the Global Thermohaline Circulation', *Paleoceanography*, v. 7, pp. 529-541.

Stocker, T.F., and Schmittner, A. (1997): 'Influence of CO_2 Emission Rates on the Stability of the Thermohaline Circulation', *Nature*, v. 388, pp. 862-865.

Taylor K.C., Lamorey, G.W., Doyle, G.A., Alley, R.B., Grootes, P.M., Mayewski, P.A., White, J.W.D., and Barlow, L.K. (1993): 'The "Flickering Switch" of Late Pleistocene Climate Change', *Nature*, v. 361, pp. 432-436.

Teller, J.T., and Thorleifson, L.H. (1983): 'The Lake Agassiz-Lake Superior Connection', in Teller, J.T. and Clayton, L. (eds.), *Glacial Lake Agassiz: Geological Association of Canada Special Paper 26*, pp. 261-290.

Thompson, L.G., Davis, M.E., Mosley-Thompson, E., Sowers, T.A., Henderson, K.A., Zagorodnov, V.S., Lin, P.-N., Mikhalenko, V.N., Campen, R.K., Bolzan, J.F., Cole-Dai, J., and Francou, B. (1998): 'A 25,000-year Tropical Climate History from Bolivian Ice Cores', *Science*, v. 282, pp. 1858-1864.

II.

PROBLEMS OF MANKIND'S SURVIVAL: RESPONSE TO THE THREAT OF CATASTROPHES WHICH CAN HAPPEN AT ANY MOMENT

IS OUR CLIMATE STABLE? BIFURCATIONS, TRANSITIONS AND OSCILLATIONS IN CLIMATE DYNAMICS

MICHAEL GHIL

1. Introduction and Motivation

Humanity has had for centuries a disruptive, as well as beneficial, effect on its local environment: urban pollution and changes in flora and fauna over large fractions of the Earth's surface go back to the rise of major civilizations in Africa, the Americas, Asia and Europe several millennia ago. Increasing industrialization and the spread of industrial methods to agriculture, forestry and fisheries at the end of the second millennium raise the possibility of the disruptive effects becoming global in the next centuries. In order to assess whether, to what extent, and in which ways, we are modifying our global environment, it is essential to understand how this environment functions. We take therefore a planetary view of the Earth's climate system, of the pieces it is made of, and of the way these pieces interact. This will allow us to understand how we might be acting on the individual pieces, and thus on the whole.

The global climate system is composed of a number of subsystems – atmosphere, biosphere, cryosphere, hydrosphere and lithosphere – each of which has distinct characteristic times, from days and weeks to centuries and millennia (Ghil and Childress, 1987; Trenberth, 1992). The atmosphere has a characteristic time of days-to-weeks in terms of the life cycle of extratropical weather systems, and of months-to-years in terms of the global mixing of trace gases. The oceans have a characteristic time as short as months-to-years in terms of their counterpart of weather systems, the life cycles of meanders and rings of the major wind-driven ocean currents. The oceans' characteristic time is as long as centuries-to-millennia in terms of their global mixing through the overturning circulation that is driven by

temperature and salinity contrasts. Snow cover and sea ice have a huge seasonal cycle, as well as sub-and interannual variability, while continental ice sheets take many millennia to build up and at least centuries to collapse.

Each subsystem has therewith its own internal variability, all other things being constant, over a fairly broad range of time scales. These ranges overlap between subsystems, due to the interactions between the latter, and thus give rise to climate-system variability on all time scales.

Can we hope to predict with confidence, and eventually control in a rational way, the effects of human intervention in this complex system? For humankind to survive and develop in a sustainable way in such a complex global environment as the climate system, it must at least understand the most basic workings of this system. We outline here the rudiments of the way in which dynamical systems theory is starting to provide such an understanding.

In Section 2, we describe the climate system's dominant balance between incoming solar radiation and outgoing terrestrial radiation. This balance is consistent with the existence of multiple equilibria of surface temperatures (Held and Suarez, 1974; Ghil, 1976; North *et al.*, 1981). Such multiple equilibria are also present for other balances of climatic actions and reactions, like the thermal driving of the mid-latitude westerly winds' being countered by surface friction and mountain drag (Charney and De Vore, 1979; Ghil *et al.*, 1991). These multiple equilibria typically arise from saddle-node bifurcations of the governing equations (Ghil, 1994). Transitions from one equilibrium to another may result from small and random pushes, a typical case of minute causes having large effects in the long term.

In Section 3, we describe the ocean's overturning circulation between cold regions where water is heavier and sinks and warm regions where it is lighter and rises. The just described effect of temperature on the water masses' density and hence motion is in competition with the effect of salinity, and hence density, increases through evaporation and brine formation *vs.* decreases in salinity and density through precipitation and river run-off. These competing effects can also give rise to two distinct equilibria (Stommel, 1961; Marotzke *et al.*, 1988). In the present-day oceans, a *thermohaline* circulation prevails, in which the temperature effects dominate. In the remote past, about 50 Myr ago, a *halothermal* circulation appears to have obtained, with salinity effects dominating (Broecker *et al.*, 1985; Kennett and Stott, 1991). In a simplified mathematical setting, these two equilibria arise by a pitchfork bifurcation that breaks the problem's mirror symmetry (Quon and Ghil, 1992; Thual and McWilliams, 1992).

On shorter time scales, of decades-to-millennia (Martinson *et al.*, 1995), oscillations of intensity and spatial pattern in the thermohaline cir-

culation seem to be the dominant mode of variability (Chen and Ghil, 1995). We show how interdecadal oscillations in the ocean's circulation arise by Hopf bifurcation (Quon and Ghil, 1995; Chen and Ghil, 1996).

In Section 4, we discuss the implications of multiple equilibria and interdecadal oscillations for our understanding of the effects that human activities might have on the climate system. The system's predictability in the absence of such effects is presented (Lorenz, 1963a, 1969; Ghil *et al.*, 1985, 1991; Ghil and Jiang, 1998). Tentative conclusions are drawn about the identification and optimization of human effects on the climate system.

2. Energy-Balance Models and the Modeling Hierarchy

2.1. *Climate Dynamics and the Global Environment*

A view of climate dynamics as a modern scientific discipline first emerged about 40 years ago (Pfeffer, 1960). We understand it at this turn of the century as studying the variability of the atmosphere-ocean-cryosphere-biosphere-lithosphere system on time scales longer than the life span of individual weather systems and shorter than the age of our planet.

When defined in these broad terms, the variability of the climate system is characterized by a power spectrum that has three components. The first is a "warm-colored" broad-band component, with power increasing from high to low frequencies. The second is a line component associated with purely periodic forcing, annual and diurnal. The third represents a number of broad peaks that might arise from less purely periodic forcing (*e.g.*, orbital change or solar variability), internal oscillations, or a combination of the two (Mitchell, 1976; Ghil and Childress, 1987, Ch. 11; Ghil and Le Treut, 1999).

Understanding the climatic mechanism or mechanisms that give rise to a particular broad peak or set of peaks represents a fundamental problem of climate dynamics. The regularities are of interest in and of themselves, for the order they create in our sparse and inaccurate observations; they also facilitate prediction for time intervals comparable to the periods associated with the given regularity (Ghil and Childress, 1987, Sec. 12.6; Ghil and Jiang, 1998).

The climate system is highly complex, its main subsystems have very different characteristic times, and the specific phenomena involved in each one of the climate problems defined in the preceding paragraphs are quite diverse. It is inconceivable, therefore, that a single model could successfully be used to incorporate all the subsystems, capture all the phenomena, and

solve all the problems. Hence the concept of a hierarchy of climate models, from the simple to the complex, has been developed about a quarter of a century ago (Schneider and Dickinson, 1974).

2.2. *Radiation Balance and Energy-Balance Models (EBMs)*

At present, the best-developed hierarchy is for atmospheric models; we summarize this hierarchy following Ghil and Robertson (2000). The first rung is formed by zero-dimensional (0-D) models, where the number of dimensions, from zero to three, refers to the number of independent space variables used to describe the model domain, *i.e.* to physical-space dimensions. Such 0-D models essentially attempt to follow the evolution of global surface-air temperature $\overline{T}$ as a result of changes in global radiative balance (Crafoord and Källén, 1978; Ghil and Childress, 1987, Sec. 10.2):

$$c\,\frac{d\overline{T}}{dt} = R_i - R_o\,, \tag{2.1a}$$

$$R_i = \mu Q_0\{1 - \alpha(\overline{T})\}, \quad R_o = \sigma m(\overline{T})\overline{T}^4. \tag{2.1b, c}$$

Here R_i and R_o are incoming solar radiation and outgoing terrestrial radiation. The heat capacity c is that of the global atmosphere, plus that of the global ocean or some fraction thereof, depending on the time scale of interest: one might only include in c the ocean mixed layer when interested in subannual time scales but the entire ocean when studying paleoclimate. The rate of change of $\overline{T}$ with time t is given by $d\overline{T}/dt$, while Q_0 is the solar radiation received at the top of the atmosphere, σ is the Stefan-Boltzmann constant, and, μ is an insolation parameter, equal to unity for present-day conditions. To have a closed, self-consistent model, the planetary reflectivity or albedo α and greyness factor m have to be expressed as functions of $\overline{T}$; $m = 1$ for a perfectly black body and $0 < m < 1$ for a grey body like planet Earth.

There are two kinds of one-dimensional (1-D) atmospheric models, for which the single spatial variable is latitude or height, respectively. The former are so-called *energy-balance models* (EBMs: Budyko, 1969; Sellers, 1969), which consider the generalization of the model (2.1) for the evolution of surface-air temperature $T = T(x, t)$, say,

$$c(x)\,\frac{\partial T}{\partial t} = R_i - R_o + D\,. \tag{2.2}$$

Here the terms on the right-hand side can be functions of the meridional coordinate x (latitude, co-latitude, or sine of latitude), as well as of time t and temperature T. The horizontal heat-flux term D expresses heat exchange between latitude belts; it typically contains first and second partial derivatives of T with respect to x. Hence the rate of change of local temperature T with respect to time also becomes a partial derivative, $\partial T/\partial t$.

The first striking results of theoretical climate dynamics were obtained in showing that slightly different forms of Eq. (2.2) could have two stable steady-state solutions, depending on the value of the insolation parameter μ [see Eq. (2.1b)] (Held and Suarez, 1974; Ghil, 1976; North *et al.*, 1981). In its simplest form, this multiplicity of stable steady states, or physically possible "climates" of our planet, can be explained in the 0-D model (2.1) by the fact that – for a fairly broad range of μ-values around $\mu = 1.0$ – the curves for R_i and R_o as a function of $\overline{T}$ intersect in 3 points. One of these corresponds to the present climate (highest $\overline{T}$-value), and another one to an ice-covered planet (lowest $\overline{T}$-value); both of these are stable, while the third one (intermediate $\overline{T}$-value) is unstable. To obtain this result, it suffices to assume that $\alpha = \alpha(\overline{T})$ is a piecewise-linear function of $\overline{T}$, with high albedo at low temperature, due to the presence of snow and ice, and low albedo at high $\overline{T}$, due to their absence, while $m = m(\overline{T})$ is a smooth, increasing function of $\overline{T}$ that attempts to capture in its simplest from the "greenhouse effect" of trace gases and water vapor (Ghil and Childress, 1987, Ch. 10).

The *bifurcation diagram* of such a 1-D EBM is shown in fig. 1. It displays the model's mean temperature $\overline{T}$ as a function of the fractional change μ in the insolation Q at the top of the atmosphere. The 'S'-shaped curve in the figure arises from two back-to-back saddle-node bifurcations. The normal form of the first one is

$$\dot{X} = \mu - X^2 . \tag{2.3a}$$

Here X stands for a suitably normalized form of T, $\dot{X} \equiv dX/dt$ and μ is a parameter that measures the stress on the system, in particular a normalized form of the insolation parameter.

The upper-most branch corresponds to the steady-state solution $X = +\sqrt{\mu}$ of Eq. (2.3a) and is stable. It matches rather well Earth's present-day climate for $\mu = 1.0$, more precisely the steady-state solution $T = T(x; \mu)$ of the full 1-D EBM (not shown) matches closely the annual mean temperature profile from instrumental data over the last century (Ghil, 1976).

The intermediate branch starts out at the left as the second solution,

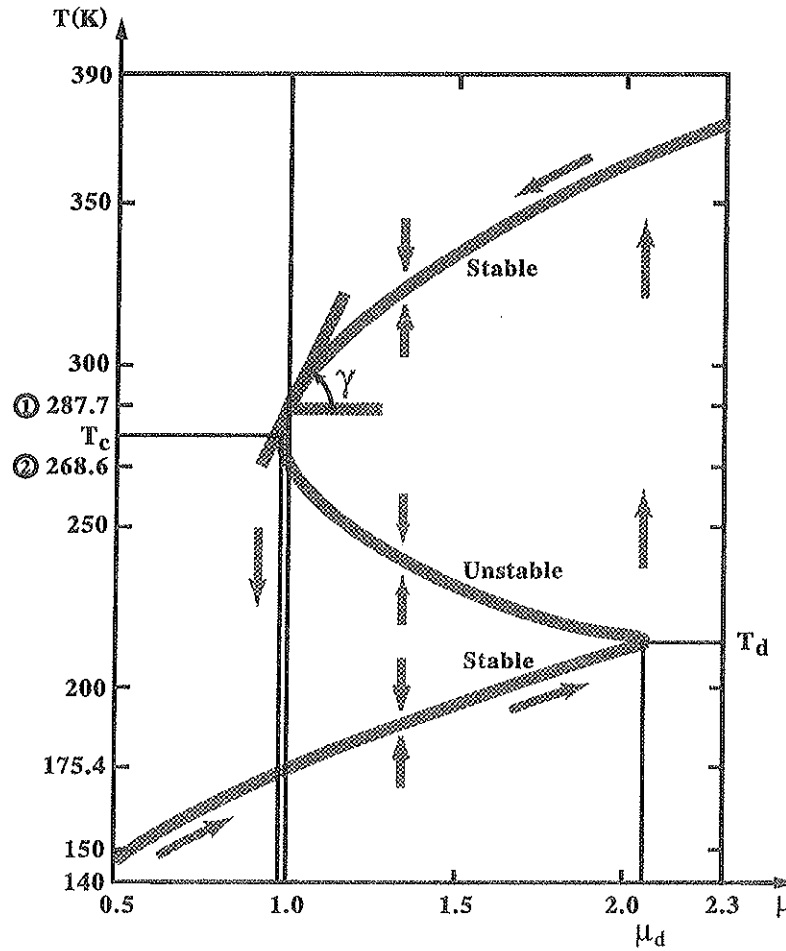


Fig. 1. Bifurcation diagram for the solutions of an energy-balance model (EBM). Annual-mean temperature $\bar{T}$ *vs.* fractional change of insolation at the top of the atmosphere μ. The arrows pointing up and down at about $\mu = 1.4$ indicate the stability of the branches: towards a given branch if it is stable and away if it is unstable. The other arrows show the hysteresis cycle that global temperatures would have to undergo for transition from the upper stable branch to the lower one and back. The angle γ gives the measure of the present climate's sensitivity to changes in insolation (after Ghil and Childress, 1987).

$X = -\sqrt{\mu}$, of Eq. (2.3a) and is unstable. It blends smoothly into the upper branch of a coordinate-shifted and mirror-reflected version of (2.3a) say

$$X' = \mu - \mu_0 + (X - X_0)^2. \tag{2.3b}$$

This branch, $X = X_0 + \sqrt{\mu_0 - \mu}$, is also unstable.

Finally, the lowermost branch in fig. 1 is the second steady-state solution of Eq. (2.3b), $X = X_0 - \sqrt{\mu_0 - \mu}$, and is also stable. It corresponds to an ice-covered planet at the same distance from the Sun as Earth.

The fact that the upper-left bifurcation point in fig. 1 is so close to present-day insolation values created great concern in the climate dynamics community in the mid-1970s, when these results were obtained. Indeed, much more detailed computations (see Sec. 2.3 below) confirmed that a reduction of about 2–5% of insolation values created great concern in the climate dynamics community in the mid-1970s, when these results were obtained. Indeed, much more detailed computations (see Sec. 2.3 below) confirmed that a reduction of about 2–5% of insolation would suffice to precipitate Earth into a "deep freeze" (Wetherald and Manabe, 1975). The great distance of the lower-right bifurcation point from present-day insolation values, on the other hand, suggests that one would have to nearly double atmospheric opacity, say, for the Earth's climate to jump back to more comfortable temperatures.

2.3. *Other Atmospheric Processes and Models*

The 1-D atmospheric models in which the details of radiative equilibrium are investigated with respect to a height coordinate z (geometric height, pressure, etc.) are often called *radiative-convective models* (Manabe and Strickler, 1964; Ramanathan and Coakley, 1978; Charlock and Sellers, 1980). This name emphasizes the key role that convection plays in vertical heat transfer. While these models preceded historically EBMs as rungs on the modeling hierarchy, it was only recently shown that they, too, can exhibit multiple equilibria (Li *et al.*, 1997; Rennó, 1997). The word (stable) "equilibrium", here and in the rest of this chapter, refers simply to a (stable) steady state of the model, rather than to true thermodynamic equilibrium.

Two-dimensional atmospheric models are also of two kinds, according to the third space coordinate which is not explicitly included. Models that resolve explicitly two horizontal coordinates, on the sphere or on a plane tangent to it, tend to emphasize the study of the dynamics of large-scale

atmospheric motions (see Sec. II in Ghil and Robertson, 2000). They often have a single layer (Charney and DeVore, 1979; Legras and Ghil, 1985) or two (Lorenz, 1963b; Reinhold and Pierrehumbert, 1982). Those that resolve explicitly a meridional coordinate and height are essentially combinations of EBMs and radiative-convective models and emphasize therewith the thermodynamic state of the system, rather than its dynamics (Saltzman and Vernekar, 1972; MacCracken and Ghan, 1988; Gallée *et al.*, 1991; Berger *et al.*, 1998). Yet another class of "horizontal" 2-D models is the extension of EBMs to resolve zonal, as well as meridional surface features, in particular land-sea contrasts (Adem, 1970; North *et al.*, 1983; Chen and Ghil, 1996).

Additional types of 1-D and 2-D atmospheric models are discussed and references to these and to the types discussed above are given by Schneider and Dickinson (1974) and by Ghil and Robertson (2000), along with some of their main applications. Finally, to encompass and resolve the main atmospheric phenomena with respect to all three spatial coordinates, *general circulation models* (GCMs) occupy the pinnacle of the modeling hierarchy (Randall, 2000).

Ghil and Robertson (2000) also describe the separate hierarchies that have grown over the last quarter-century in modeling the ocean and the coupled ocean-atmosphere system. More recently, an overarching hierarchy of earth-system models – that encompass all the subsystems of interest, atmosphere, biosphere, cryosphere, hydrosphere and lithosphere – has been developing. Eventually, the partial results about each subsystem's variability, outlined in this section and the next one, will have to be verified from one rung to the next of the earth-system modeling hierarchy.

In the meantime, it is worth noting that the results of climate simulations with GCMs, whether atmospheric or coupled, are often still interpreted in terms of the understanding gained from 0-D or 1-D EBMs. Wetherald and Manabe (1975), using a simplified sectorial GCM, confirmed the dependence of mean zonal temperature on the insolation parameter μ (the normalized "solar constant") obtained for 1-D EBMs (see Ghil and Childress, 1987, Ch. 10). These authors could not confirm the presence of the intermediate, unstable branch or of the "deep-freeze" stable branch in fig. 1 with their GCM, because of the model's computational limitations. But the parabolic shape of the upper, present-day-like branch near the upper-left bifurcation point in our figure [cf. Eq. (2.3a)] was well supported by their GCM simulations.

In fact, the sensitivity $\tan \gamma = (d\overline{T}/d\mu) \mid_{\mu=1.0}$ of global temperature $\overline{T}$ to changes in μ near the present-day climate (see fig. 1) equals about 1K per 1% change in the insolation for both EBMs and GCMs. Many GCM studies of climate-change response to increases in greenhouse trace-gas concentrations use therefore even a linearized version of Eq. (2.1),

$$c\frac{d\overline{T}}{dt} = \lambda T + Q\,, \tag{2.4a}$$

$$\lambda = \sum_{i=1}^{I} \lambda_i\,, \qquad Q = \sum_{j=1}^{J} Q_j\,, \tag{2.4b,c}$$

for interpreting the roles of the different feedbacks λ_i, positive ($\lambda_i < 0$) or negative ($\lambda_i > 0$), and heat sources, $Q_j > 0$, or sinks, $Q_j < 0$ (*e.g.*, Schlesinger and Mitchell, 1987; Cess, Potter, *et al.*, 1989; Li and Le Treut, 1992).

3. Interdecadal Oscillations in the Oceans' Thermohaline Circulation

3.1. *Theory and Simple Models*

Historically, the thermohaline circulation (THC) was first among the climate system's major processes to be studied using a very simple mathematical model and be shown to possess multiple equilibria (Stommel, 1961). A sketch of the Atlantic Ocean's THC and its interactions with the atmosphere and cryosphere on long time scales is shown in fig. 2. These interactions can lead to climate oscillations with multi-millennial periods – such as the Heinrich events (Ghil, 1994, and references therein) – and are summarized in the figure's caption, following Ghil *et al.* (1987). An equally schematic view of the global THC is provided by the widely known "conveyor belt" diagram (*e.g.*, Broecker, 1991), which does not commonly include these interactions.

Basically, the THC is due to denser water sinking, lighter water rising, and water-mass continuity closing the circuit through near-horizontal flow between the areas of rising and sinking. This is roughly the oceanic equivalent of the atmosphere's Hadley circulation, with two notable differences:

i) The ocean water's density ϱ is a function of temperature T and salinity S, while that of the air depends on temperature and humidity.

ii) Water sinks in and near fairly concentrated regions of intense convection, currently located mostly in high latitudes, and rises diffusely over the rest of the ocean. Air, on the other hand, does rise most intensely in cumulus towers, but overall the areas of net rising and sinking air in a Hadley cell are quite comparable in extent, when viewed on the synoptic and planetary scales.

The effects of temperature and salinity on the ocean water's density, $\varrho = \varrho(T, S)$, oppose each other: the density ϱ *decreases* with increasing T and *increases* with increasing S. It is these two effects that give the *thermohaline*

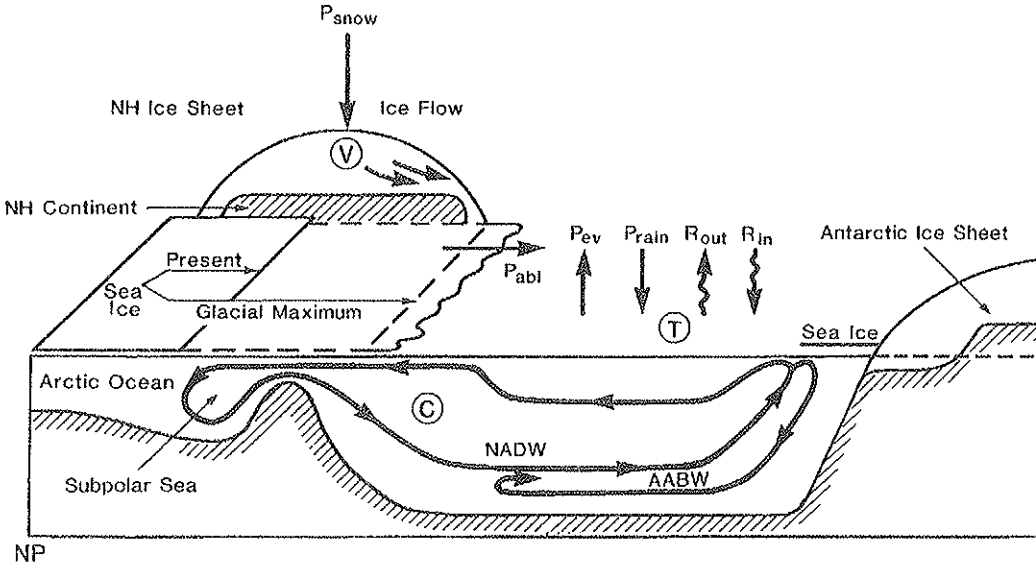


Fig. 2. Diagram of an Atlantic meridional cross section from North Pole (NP) to South Pole (SP), showing mechanisms likely to affect the thermohaline circulation (THC) on various time-scales. Changes in the radiation balance R_{in}-R_{out} are due, at least in part, to changes in extent of Northern Hemisphere (NH) snow and ice cover, V, and how they affect the global temperature, T; the extent of Southern Hemisphere ice is assumed constant, to a first approximation. The change in hydrologic cycle expressed in the terms P_{rain}-P_{evap} for the ocean and P_{snow}-P_{abl} for the snow and ice is due to changes in ocean temperature. Deep-water formation in the North Atlantic Subpolar Sea (North Atlantic Deep Water: NADW) is affected by changes in ice volume and extent, and regulates the intensity C of the THC; changes in Antarctic Bottom Water (AABW) formation are neglected in this approximation. This in turn affects the system's temperature, and is also affected by it (after Ghil *et al.*, 1987).

circulation its name, from the Greek words for *T* and *S*. In high latitudes, ϱ increases as the water loses heat to the air above and, if sea ice is formed, as the water underneath is enriched in brine. In low latitudes, ϱ increases due to evaporation but decreases due to heat flux into the ocean.

For the present climate, the temperature effect is commonly assumed to be stronger than the salinity effect, and ocean water is observed to sink in certain areas of the high-latitude North Atlantic and Southern Ocean – with very few and limited areas of deep-water formation elsewhere – and to rise everywhere else. Thus *thermohaline*, *T* more important than and hence before *S*. During some remote geological times, deep water apparently formed in the global ocean near the equator; such an overturning circulation of opposite sign to that prevailing today has been dubbed *halothermal*,

S before T (*e.g.*, Kennett and Stott, 1991). The quantification of the relative effects of T and S on the oceanic water masses' buoyancy in high and low latitudes is far from complete, especially for paleocirculations; the association of the latter with salinity effects that exceed the thermal ones is thus rather tentative.

Stommel (1961) considered a two-box model, with two pipes connecting the two boxes. He showed that the system of two nonlinear, coupled ordinary differential equations (ODEs) which govern the temperature and salinity differences between the two well-mixed boxes has two stable steady-state solutions, distinguished by the direction of flow in the upper and lower pipe. Stommel's paper was primarily concerned with distinct local convection regimes, and hence vertical stratifications, in the North Atlantic and Mediterranean (or Red Sea), say. Today, we mainly think of one box as representing the low latitudes and the other one the high latitudes in the global THC.

The next step in the hierarchical modeling of the THC is that of 2-D meridional plane models (see Sec. I.B of Ghil and Robertson, 2000), in which the temperature and salinity fields are governed by coupled nonlinear partial differential equations with two independent space variables, latitude and depth, say. Given boundary conditions for such a model that are symmetric about the Equator, as are the equations themselves, one expects a symmetric solution, in which water either sinks near the poles and rises everywhere else (thermohaline) or sinks near the Equator and rises everywhere else (halothermal). These two symmetric solutions would correspond to the two equilibria of Stommel's (1961) box model.

In fact, fig. 3 shows that symmetry breaking can occur, leading gradually from a symmetric two-cell circulation (fig. 3a) to an antisymmetric one-cell circulation (approximately achieved in fig. 3c). In between, all degrees of dominance of one cell over the other are possible, with one such intermediate state shown in fig. 3b. A situation lying somewhere between figs. 3b and 3c seems to resemble most closely the meridional overturning diagram of the Atlantic Ocean in fig. 2.

This symmetry breaking can be described by a *pitchfork bifurcation* (*e.g.*, Guckenheimer and Holmes, 1983):

$$\dot{X} = \mu - X^3 . \tag{3.1}$$

Here X stands for the amount of asymmetry in the solution, so that $X = 0$ is the symmetric branch, and μ is a parameter that measures the stress on the system, in particular a normalized form of the buoyancy flux at the

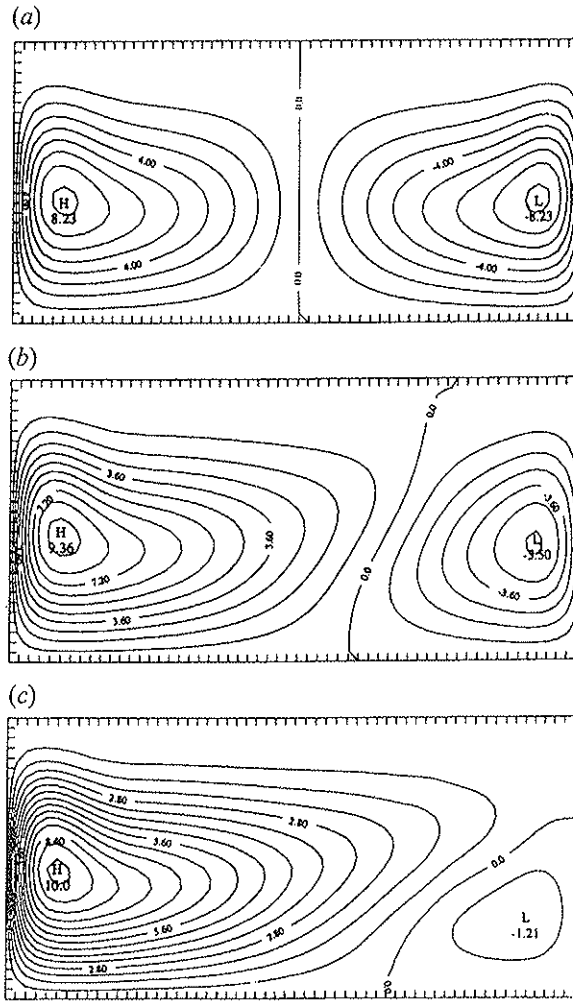


Fig. 3. Stream-function fields for a 2-D, meridional plane THC model with so-called mixed boundary conditions: the temperature profile and salinity flux are imposed at one horizontal boundary of the rectangular box, while the other three boundaries are impermeable to heat and salt. (a) Symmetric solution for low salt-flux forcing; (b, c) increasingly asymmetric solutions as the forcing is increased (from Quon and Ghil, 1992).

surface. For $\mu < 0$ the symmetric branch is stable, while for $\mu > 0$ the two branches $X = \pm\sqrt{\mu}$ inherit its stability.

Thus, figs. 3b and 3c both lie on a solution branch of the 2-D THC problem for which the left cell dominates: say that North Atlantic Deep Water extends to the Southern Ocean's polar front, as it does in fig. 2. According to Eq. (3.1), another branch exists, whose flow patterns are mirror images in the rectangular box's vertical symmetry axis (the "equatorial plane") of those in figs. 3b and 3c. The existence of this second branch was verified numerically by Quon and Ghil (1992; their fig. 16). Thual and McWilliams (1992) considered more complex bifurcation diagrams for a similar 2-D model and showed the equivalence of such a diagram for their 2-D model and a box-and-pipe model of sufficient complexity.

3.2. *Bifurcation diagrams for GCMs*

Bryan (1986) was the first to document transition from a two-cell to a one-cell circulation in a simplified ocean GCM with idealized, symmetric forcing, in agreement with the three-box scenario of Rooth (1982). Manabe and Stouffer (1999), among others, questioned however the realism of more than one stable THC equilibrium by using coupled ocean-atmosphere GCMs. The situation with respect to the THC's pitchfork bifurcation (3.1) is thus subtler than it was with respect to fig. 1 for radiative equilibrium. While in Sec. 2 atmospheric GCMs confirmed essentially the EBM results, climbing the rungs of the modeling hierarchy for the THC yields contradictory results that are still in need of further clarification.

Internal variability of the THC – with smaller and more regular excursions than the huge and totally irregular jumps associated with bistability – was studied intensively in the late 1980s and the 1990s. These studies placed themselves on various rungs of the modeling hierarchy, from Boolean delay equation models (so-called "formal conceptual models": Ghil *et al.*, 1987; Darby and Mysak, 1993) through box models (Welander, 1986) and 2-D models (Quon and Ghil, 1995) to ocean GCMs. A summary of the different kinds of oscillatory variability found in the latter appears in Table I below. Additional GCM references for these three types of oscillations are given by McWilliams (1996). Such oscillatory behavior seems to match more closely the instrumentally recorded THC variability (see Sec. 3.3 below), as well as the paleoclimatic records for the recent geological past (see Ghil, 1994), than bistability.

The interaction of the (multi-)millennial oscillations with variability in the surface features and processes shown in fig. 2 is discussed by Ghil (1994). Chen and Ghil (1996), in particular, studied some of the interactions

Table I. *Thermohaline circulation (THC) oscillations* (adapted from Ghil, 1994).

Time scale	*Phenomena*	*Mechanism*
Interdecadal	3-D, wind-driven + thermohaline circulation	- Gyre advection (Weaver *et al.*, 1991, 1993). - Localized surface-density anomalies due to surface coupling (Chen and Ghil, 1995, 1996).
Centennial	Loop-type, Atlantic-Pacific circulation	Conveyor-belt advection of high-latitude density anomalies (Mikolajewicz and Maier-Reimer, 1990).
Millennial	Relaxation oscillation, with "flushes" and superimposed decadal fluctuations	Bottom-water warming, due to high-latitude freshening and its braking effect (Marotzke, 1989; Chen and Ghil, 1995).

between atmospheric processes and the THC. They used a so-called hybrid coupled model, to wit a (horizontally) 2-D EBM (see Sec. 2.3) coupled to a rectangular-box version of the North Atlantic rendered by a low-resolution ocean GCM. This hybrid model's regime diagram is shown in fig. 4a. A steady state is stable for high values of the coupling parameter λ_{ao} or of the EBM's diffusion parameter d. Interdecadal oscillations with a period of 40-50 years are self-sustained and stable for low values of these parameters.

Chen and Ghil (1996) studied some of the interactions between atmospheric processes and the THC. They used a so-called hybrid coupled model, to wit a (horizontally) 2-D EBM (see Sec. 2.3) coupled to a rectangular-box version of the North Atlantic rendered by a low-resolution ocean GCM. This hybrid model's regime diagram is shown in fig. 4a. A steady state is stable for high values of the coupling parameter λ_{ao} or of the EBM's diffusion parameter d. Interdecadal oscillations with a period of 40-50 years are self-sustained and stable for low values of these parameters.

The self-sustained THC oscillations in question are characterized by a pair of vortices of opposite sign that grow and decay in quadrature with each other in the ocean's upper layers. Their centers follow each other anti-clockwise through the northwestern quadrant of the model's rectangular

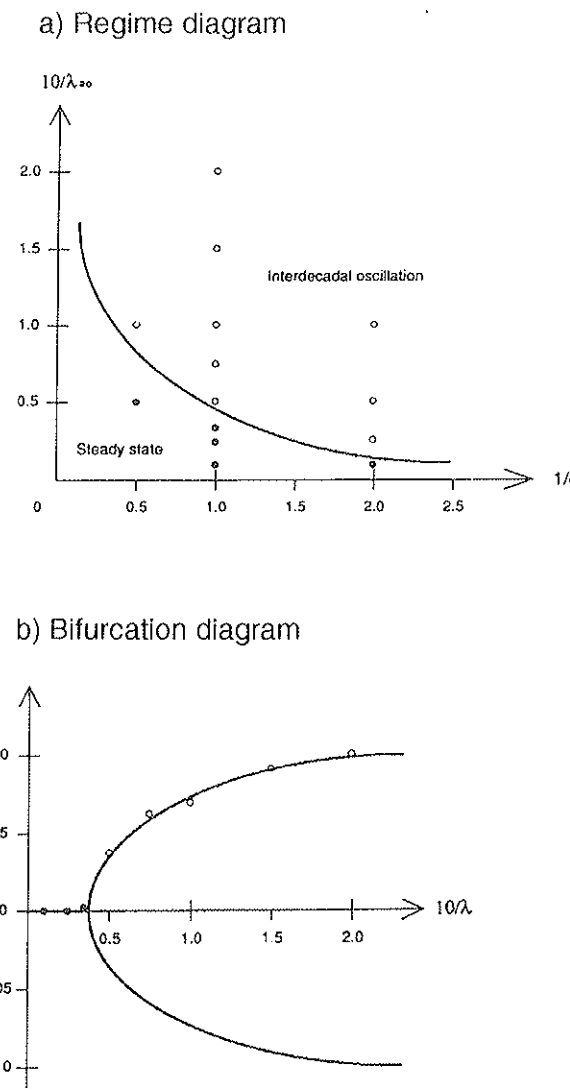


Fig. 4. Dependence of THC solutions on two parameters in a hybrid coupled model (HCM); the two parameters are the atmosphere-ocean coupling coefficient λ_{ao} and the atmospheric thermal diffusion coefficient d. (a) Schematic regime diagram. The full circles stand for the model's stable steady states, the open circles for stable limit cycles, and the solid curve is the estimated neutral stability curve between the former and the latter. (b) Hopf bifurcation curve at fixed $d = 1.0$ and varying λ_{ao}; this curve was obtained by fitting a parabola to the model's numerical-simulation results, shown as full and open circles (from Chen and Ghil, 1996).

domain. Both the period and the spatio-temporal characteristics of the oscillation are thus rather similar to those seen in a fully coupled GCM with realistic geometry (Delworth *et al.*, 1993).

The transition from a stable equilibrium to a stable limit cycle, via Hopf bifurcation, in Chen and Ghil's hybrid coupled model is shown in fig. 4b. The physical characteristics of the oscillatory instability that leads to the Hopf bifurcations have been described in further detail by Colin de Verdière and Huck (1999), using both a four-box ocean-atmosphere and a number of more detailed models.

3.3. *Interannual and Interdecadal Climate Variability*

Human intervention in the workings of climate on the global scale is most likely on time scales comparable to those of major socio-economic changes. The latter occur typically, at this turn of the century, on interannual to interdecadal time scales.

The best-known climatic regularities on the interannual time scale are the quasi-biennial and low-frequency (4-5-year) component of the El-Niño/Southern-Oscillation (ENSO) phenomenon (Rasmusson *et al.*, 1990; Neelin *et al.*, 1998). A particularly appealing explanation of these two spectral peaks that is also consistent with'the broad-band spectrum in the 1-10-year band is given by the Devil's staircase mechanism (Jin *et al.*, 1994, 1996; Tziperman *et al.*, 1994).

This mechanism is most easily explained in two steps. First, it is by now well accepted that the coupled ocean-atmosphere system in the tropical Pacific exhibits a self-sustained oscillation with a period of 2-3 years for mean-annual conditions. Second, this self-sustained oscillation interacts with the seasonal cycle in insolation and frequency-locks preferentially to an integer multiple of the period in the forcing: 2, 3, 4 or 5 years. Jumps between these broad integer steps in the Devil's staircase – as well as between narrower, rational steps – occur as the result of atmospheric, higher-frequency "noise". This mechanism is described further and compared with other scenarios of interannual variability and with the observational evidence in Sec. III of Ghil and Robertson (2000).

Ghil and Vautard (1991) described statistically significant regularities in the interdecadal band of climatic variability. Broad peaks of roughly 13-15 and 23-27 years have been since confirmed in various records, both instrumental (Plaut *et al.*, 1995; Moron *et al.*, 1998) and paleoclimatic. The interdecadal oscillations in the THC reviewed in Sec. 3.2. above are a plausible mechanism for these peaks, but not yet generally accepted as such.

4. Concluding Remarks

Is the impact of human activities on climate observable and identifiable in the instrumental records of the last century-and-a-half? This depends on the null hypothesis against which such an impact is tested.

The current approach that is generally pursued assumes essentially that past climate variability is indistinguishable from a stochastic red-noise process (Hasselmann, 1976), whose only regularities are those of periodic external forcing (Mitchell, 1976). Given such a null hypothesis, the official consensus of IPCC (1995) tilts towards a global warming effect of recent trace-gas emissions, which exceeds the cooling effect of anthropogenic aerosol emissions.

Atmospheric and coupled GCM simulations of the trace-gas warming and aerosol cooling buttress this IPCC consensus. The GCM simulations used so far do not, however, exhibit the observed interdecadal regularities described at the end of Sec. 3.3. here. They might, therewith, miss some important physical mechanisms of climate variability and are therefore not entirely conclusive.

As Northern Hemisphere temperatures were falling in the 1960s and early 1970s, the aerosol effect was the one that caused the greatest concern. As shown in Sec. 2.2. here, this concern was bolstered by the possibility of a huge, highly nonlinear temperature drop if the climate system reached the upper-left bifurcation point of fig. 1.

The global temperature increase through the 1990s is certainly rather unusual, in terms of the instrumental record of the last 150 years or so. It does not correspond, however, to a rapidly accelerating increase in greenhouse-gas emissions or a substantial drop in aerosol emissions. How statistically significant is, therefore, this temperature rise, if the null hypothesis is not a random coincidence of small stochastic excursions of global temperatures being all, or nearly all, of the same sign?

The presence of internally arising regularities in the climate system, with periods of years and decades, suggests the need for a different null hypothesis. Essentially, the behavior of the climatic signal has to be shown distinct from that generated by natural climate variability in the past, when human effects were negligible, at least on the global scale. As discussed in Secs. 2.1. and 3.3. here, this natural variability includes interannual and interdecadal peaks, as well as the broad-band component.

Ghil and Jiang (1998) showed that – on the seasonal-to-interannual time scale – climate predictability is greatly enhanced by the presence of these regularities. This is the case not only against the classical benchmark of a red-noise process of first-order autoregressive type but also against a

purely chaotic, albeit deterministic process such as that of Lorenz (1963a).

The conclusion for the interdecadal time scale on which the global warming problem is being asked is two-fold. First, we need to describe and understand climate variability on this time scale – in particular its regularities – as well as on the interannual time scale. Once that is done, prediction with known error bars on the interdecadal time scale will be possible (see Plaut *et al.*, 1995).

With these results in hand, we should be able to proceed to the second conclusion. To wit, deviations of the current record from that predicted based on past natural variability has to be attributed to new causes. The "suspects" clearly include human effects, and attribution to them will become therewith both easier and more reliable.

At the same time, it is to be hoped that the applications of dynamical systems theory to the global socio-economic system and to populations dynamics will have made considerable progress. That being so, a rational approach to the prediction and control of the coupled system formed by all living beings on this planet and their physico-chemical environment might be possible.

Acknowledgements

It is a pleasure to thank all the students and colleagues from whom I have learned about climate dynamics and dynamical systems theory. Among the latter, A.W. Robertson has been particularly helpful in preparing this chapter. Discussions with the colleagues from many other areas of the sciences and of science policy met at the Pontifical Academy's Study Week have been most stimulating in the chapter's final revision. I am grateful for the sabbatical hospitality of the Ecole Normale Supérieure in Paris and its Département Terre-Atmosphère-Océan, as well as to the Laboratoire de Météorologie Dynamique du Centre National de la Recherche Scientifique (CNRS) at the Ecole. The secretarial support of Ms. F. Fleuriau is much appreciated, as always. My work on the topics reviewed here has been supported by an NSF Special Creativity Award.

REFERENCES

Adem, J. (1970): 'Incorporation of Advection of Heat by Mean Winds and by Ocean Currents in a Thermodynamic Model for Long-range Weather Prediction', *Mon. Weather Rev.*, *98*, pp. 776-786.

Berger, A., Loutre, M.F., and Gallée, H. (1998): 'Sensitivity of the LLN Climate Model to the Astronomical and CO_2 Forcings over the Last 200 kyr', *Clim. Dyn.*, *14*, pp. 615-629.

Broecker, W.S., Peteet, D.M., and Rind, D. (1985): 'Does the Ocean-atmosphere System have more than one Stable Mode of Operation?', *Nature*, *315*, pp. 21-25.

Broecker, W.S. (1991): 'The Great Ocean Conveyor', *Oceanography*, *4*, pp. 79-89.

Budyko, M.I. (1969): 'The Effect of Solar Radiation Variations on the Climate of the Earth', *Tellus*, *21*, pp. 611-619.

Cess, R.D., Potter, G.L., Blanchet, J.P., Boer, G.J., Ghan, S.J., Kiehl, J.T., Le Treut, H., Li, Z.-X., Liang, X.-Z., Mitchell, J.F.B., Morcrette, J.-J., Randall, D.A., Riches, M.R., Roeckner, E., Schlese, U., Slingo, A., Taylor, K.E., Washington, W.M., Wetherald, R.T., and Yagai, I. (1989): 'Interpretation of Cloud-climate Feedbacks as Produced by 14 Atmospheric General Circulation Models', *Science*, *245*, pp. 513-551.

Charney, J.G., and DeVore, J.G. (1979): 'Multiple Flow Equilibria in the Atmosphere and Blocking', *J. Atmos. Sci.*, *36*, pp. 1205-1216.

Chen, F., and Ghil, M. (1995): 'Interdecadal Variability of the Thermohaline Circulation and High-latitude Surface Fluxes', *J. Phys. Oceanogr.*, *25*, pp. 2547-2568.

Chen, F., and Ghil, M. (1996): 'Interdecadal Variability in a Hybrid Coupled Ocean-atmosphere Model', *J. Phys. Oceanogr.*, *26*, pp. 1561-1578.

Colin de Verdière, A., and Huck, T. (1999): 'Baroclinic Instability: An oceanic Wavemaker for Interdecadal Variability', *J. Phys. Ocean.*, *29*, pp. 893-910.

Crafoord, C., and Källen, E. (1978): 'A Note on the Condition for Existence of more than one Steady-state Solution in Budyko-Sellers Type Models', *J. Atmos. Sci.*, *35*, pp. 1123-1125.

Darby M.S., and Mysak, L.A. (1993): 'A Boolean Delay Equation Model of an Interdecadal Arctic Climate Cycle', *Clim. Dyn.*, *8*, pp. 241-246.

Delworth, T.S., Manabe, S., and Stouffer, R.J. (1993): 'Interdecadal Variations of the Thermohaline Circulation in a Coupled Ocean-atmosphere Model', *J. Climate*, *6*, pp. 1993-2011.

Gallée, H., van Ypersele, J.P., Fichefet, Th., Tricot, C., and Berger, A. (1991): 'Simulation of the Last Glacial Cycle by a Coupled, Sectorially, Averaged Climate – Ice-sheet Model. I. The Climate Model', *J. Geophys. Res.*, *96*, 13, pp. 139-161.

Ghil, M. (1976): 'Climate Stability for a Sellers-type Model', *J. Atmos. Sci.*, *33*, pp. 3-20.

Ghil, M., and Le Treut, H. (1981): 'A Climate Model with Cryodynamics and Geodynamics', *J. Geophys. Res.*, *86*, pp. 5262-5270.

Ghil, M., Benzi, R., and Parisi, G. (eds.) (1985): *Turbulence and Predictability in Geophysical Fluid Dynamics and Climate Dynamics* (North-Holland, Amsterdam/NewYork/Oxford/Tokyo), 449 pp.

Ghil, M., and Childress, S. (1987): *Topics in Geophysical Fluid Dynamics: Atmospheric Dynamics, Dynamo Theory and Climate Dynamics* (Springer-Verlag, New York/Berlin/London/Paris/Tokyo), 485 pp.

Ghil, M., and Vautard, R. (1991): 'Interdecadal Oscillations and the Warming Trend in Global Temperature Time Series', *Nature, 350*, pp. 324-327.

Ghil, M., Kimoto, M., and Neelin, J.D. (1991): 'Nonlinear Dynamics and Predictability in the Atmospheric Sciences', *Rev. Geophys. Supplement* (U.S. Nat'l Rept. to Int'l Union of Geodesy & Geophys. 1987-1990), *36*, pp. 46-55.

Ghil, M. (1994): 'Cryothermodynamics: The Chaotic Dynamics of Paleoclimate', *Physica D, 77*, pp. 130-159.

Ghil, M., and Jiang, N. (1998): 'Recent Forecast Skill for the El-Niño/Southern Oscillation', *Geophys. Res. Lett., 25*, pp. 171-174.

Ghil, M., and Le Treut, H. (1999): 'Climate Variability and Climate Change', in *Scientific Bridges for 2000 and Beyond, a Virtual Colloquium by the Elf-Aquitaine Professors of the Académie des Sciences, Rapports de l'Académie des Sciences* (TEC&DOC, London/Paris/NewYork), pp. 105-119.

Ghil, M., and Robertson, A.W. (2000): 'Solving Problems with GCMs: General Circulation Models and their Role in the Climate Modeling Hierarchy', *General Circulation Model Development: Past, Present and Future*, D. Randall (ed.), Academic Press, in press.

Guckenheimer, J., and Holmes, P. (1983): *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields* (Springer-Verlag, New York), 453 pp.

Hasselmann, K. (1976): 'Stochastic Climate Models, Part I: Theory', *Tellus, 28*, pp. 473-485.

Held, I.M., and Suarez, M.J. (1974): 'Simple Albedo Feedback Models of the Ice Caps', *Tellus, 26*, pp. 613-629.

IPCC (Intergovernmental Panel on Climate Change) (1995): *Climate Change 1995: The Science of Global Change*, J.T. Houghton, L.G. Meira Filho, B.A. Callander, N. Harris, A. Kattenberg, and K. Maskell (eds.) (Cambridge University Press, Cambridge, UK), 572 pp.

Jin, F.-F., Neelin, J.D., and Ghil, M. (1994): 'El Niño on the Devil's Staircase: Annual Subharmonic Steps to Chaos', *Science, 264*, pp. 70-72.

Jin, F.-F., Neelin, J.D., and Ghil, M. (1996): 'El Niño/Southern Oscillation and the Annual Cycle: Subharmonic Frequency-locking and Aperiodicity', *Physica D, 98*, pp. 442-465.

Kennett, R.P., and Stott, L.D. (1991): 'Abrupt Deep-sea Warming, Paleoceanographic Changes and Benthic Extinctions at the End of the Palaeocene', *Nature, 353*, pp. 225-229.

Legras, B., and Ghil, M. (1985): 'Persistent Anomalies, Blocking and Variations in Atmospheric Predictability', *J. Atmos. Sci., 42*, pp. 433-471.

Li, Z.-X., Ide, K., Le Treut, H., and Ghil, M. (1997): 'Atmospheric Radiative Equilibria in a Simple Column Model', *Clim. Dyn., 13*, pp. 429-440.

Li, Z.-X., and Le Treut, H. (1992): 'Cloud-radiation Feedbacks in a General Circulation Model and their Dependence on Cloud Modelling Assumptions', *Clim. Dyn., 7*, pp. 133-139.

Lorenz, E.N. (1963a): 'Deterministic Nonperiodic Flow', *J. Atmos. Sci., 20*, pp. 130-141.

Lorenz, E.N. (1963b): 'The Mechanics of Vacillation', *J. Atmos. Sci., 20*, pp. 448-464.

Lorenz, E.N. (1969): 'Three Approaches to Atmospheric Predictability', *Bull. Amer. Met. Soc., 50*, pp. 345-351.

MacCracken, M.C., and Ghan, S.J. (1988): 'Design and Use of Zonally Averaged Models', in *Physically-Based Modeling and Simulation of Climate and Climatic Change*, M.E. Schlesinger (ed.) (Kluwer Acad. Publishers, Dordrech), pp. 755-803.

Manabe, S., and Strickler, R.F. (1964): 'Thermal Equilibrium of the Atmosphere with a Convective Adjustment', *J. Atmos. Sci., 21*, pp. 361-385.

Manabe, S., and Stouffer, R.J. (1999): 'Are Two Modes of Thermohaline Circulation Stable?', *Tellus, 51A*, pp. 400-411.

Marotzke, J., Welander, P., and Willebrand, J. (1988): 'Instability and Multiple Steady States in a Meridional-plane Model of the Thermohaline Circulation', *Tellus, 40A*, pp. 162-172.

Marotzke, J. (1989): 'Instabilities and Multiple Steady States of the Thermohaline Circulation', in *Ocean Circulation Models: Combining Data and Dynamics*, D.L.T. Anderson and J. Willebrand (eds.) (Kluwer Academic), pp. 501-511.

Mikolajewicz, U., and Maier-Reimer, E. (1990): 'Internal Secular Variability in an Ocean General Circulation Model', *Clim. Dyn., 4*, pp. 145-156.

Martinson, D.G., Bryan, K., Ghil, M., Hall, M.M., Karl, T.R., Sarachik, E.S., Sorooshian, S., and Talley, L.D. (eds.) (1995): *Natural Climate Variability on Decade-to-Century Time Scales* (National Academy Press, Washington, D.C.) 630 pp.

McWilliams, J.C. (1996): 'Modeling the Oceanic General Circulation', *Annu. Rev. Fluid. Mech., 28*, pp. 215-248.

Mitchell, J.M. Jr. (1976): 'An Overview of Climatic Variability and its Causal Mechanisms', *Quartern. Res., 6*, pp. 481-493.

Moron, V., Vautard, R., and Ghil, M. (1998): 'Trends, Interdecadal and Interannual Oscillations in Global Sea-surface Temperatures', *Clim. Dyn., 14*, pp. 545-569.

Neelin, J.D., Battisti, D.S., Hirst, A.C., Jin, F.-F., Wakata, Y., Yasmagata, T., and Zebiak, S.E. (1998): 'ENSO Theory', *J. Geophys. Res., 103*, 14, pp. 261-290.

North, G.R., Cahalan, R.F., and Coakley, J.A. Jr. (1981): 'Energy Balance Climate Models', *Rev. Geophys. Space Phys., 19*, pp. 91-121.

North, G.R., Mengel, J.G., and Short, D.A. (1983): 'Simple Energy Balance Model Resolving the Seasons and the Continents: Application to the Astronomical Theory of the Ice Ages', *J. Geophys. Res., 88*, pp. 6576-6586.

Plaut, G., Ghil, M., and Vautard, R. (1995): 'Interannual and Interdecadal Variability in 335 Years of Central England Temperature', *Science, 268*, pp. 710-713.

Quon, C., and Ghil, M. (1992): 'Multiple Equilibria in Thermosolutal Convection due to Salt-flux Boundary Conditions', *J. Fluid Mech., 245*, pp. 449-483.

Quon, C., and Ghil, M. (1995): 'Multiple Equilibria and Stable Oscillations in Thermosolutal Convection at Small Aspect Ratio', *J. Fluid. Mech., 291*, pp. 33-56.

Ramanathan, V., and Coakley, J.A. (1978): 'Climate Modeling Through Radiative Convective Model', *Rev. Geophys. Space Phys., 16*, pp. 465-489.

Randall, D. (ed.) (2000): *General Circulation Model Development: Past, Present and Future*, Academic Press, in press.

Rasmusson, E.M., Wang, X., and Ropelewski, C.F. (1990): 'The Biennial Component of ENSO Variability', *J. Marine Syst., 1*, pp. 71-96.

Reinhold, B.B., and Pierrehumbert, R.T. (1982): 'Dynamics of Weather Regimes: Quasi-stationary Waves and Blocking', *Mon. Wea. Rev., 110*, pp. 1105-1145.

Rennó, N.O. (1997): 'Multiple Equilibria in Radiative-convective Atmospheres', *Tellus, 49A*, pp. 423-438.

Rooth, C. (1982): 'Hydrology and Ocean Circulation', *Progr. Oceanogr.*, *11*, pp. 131-149.

Salzman, B., and Vernekar, A.D. (1972): 'Global Equilibrium Solutions for the Zonally Averaged Macroclimate', *J. Geophys. Res.*, *77*, pp. 3936-3945.

Saltzman, B. (1985): 'Paleoclimatic Modeling', in *Paleoclimate Analysis and Modeling*, A.D. Hecht (ed.) (J. Wiley, New York), pp. 341-396.

Schlesinger, M.E., and Mitchell, J.B. (1987): 'Climate Model Simulations of the Equilibrium Climatic Response to Increased Carbon Dioxide', *Rev. Geophys.*, *25*, pp. 760-798.

Schneider, S.H., and Dickinson, R.E. (1974): 'Climate Modeling', *Rev. Geophys. Space Phys.*, *25*, pp. 447-493.

Sellers, W.D. (1969): 'A Climate Model Based on the Energy Balance of the Earth-atmosphere System', *J. Appl. Meteor.*, *8*, pp. 392-400.

Stommel, H. (1961): 'Thermohaline Convection with Two Stable Regimes of Flow', *Tellus*, *13*, pp. 224-230.

Thual, O., and McWilliams, J.C. (1992): 'The Catastrophe Structure of Thermohaline Convection in a Two-dimensional Fluid Model and a Comparison with Low-order Box Models', *Geophys. Astrophys. Fluid Dyn.*, *64*, pp. 67-95.

Trenberth, K. (ed.) (1992): *Climate System Modeling* (Cambridge University Press), 788 pp.

Tziperman, E., Stone, L., Cane, M., and Jarosh, H. (1994): 'El Niño chaos: Overlapping of Resonances between the Seasonal Cycle and the Pacific Ocean-atmosphere Oscillator', *Science*, *264*, pp. 72-74.

Weaver, A.J., Sarachik, E.S., and Marotzke, J. (1991): 'Freshwater Flux Forcing of Decadal and Interdecadal Oceanic Variability', *Nature*, *353*, pp. 836-838.

Weaver, A.J., Marotzke, J., Cummings, P.F., and Sarachick, E.S. (1993): 'Stability and Variability of the Thermohaline Circulation', *J. Phys. Oceanogr.*, *23*, pp. 39-60.

Welander, P. (1986): 'Thermohaline Effects in the Ocean Circulation and Related Simple Models', in *Large-Scale Transport Processes in Oceans and Atmosphere*, Willebrand and D.L.T. Anderson (eds.) (D. Reidel), pp. 163-200.

Wetherald, R.T., and Manabe, S. (1975): 'The Effect of Changing the Solar Constant on the Climate of a General Circulation Model', *J. Atmos. Sci.*, *32*, pp. 2044-2059.

NONLINEAR DYNAMICS OF LIVING NEURONS

MIKHAIL I. RABINOVICH, PABLO VARONA and HENRY D.I. ABARBANEL

1. Introduction

What is the role of neural units for living systems? Assemblies of neurons in the simplest animals to the most complex are organized to solve problems relating to the functional behavior of an animal in an often unpredictable, certainly changing, and sometimes challenging environment. In the face of such variability they must robustly, reliably, and reproducibly differentiate among different odors and sounds, identify muscle actions to acquire and digest food, and in humans provide memory and more elusively that behavior we label as consciousness. These and other tasks are accomplished by (1) the production of rhythms or particular spatio-temporal patterns of oscillation to control the rhythmic behavior of animals, and (2) processing information received from the environment. Each of these actions require activity which is *dynamical*, evolves in time and/or space, and *nonlinear*, requires the rich structure of nonproportional response to driving forces, external or internal.

The experience which has accumulated in the study of nonlinear dynamics for the description and prediction of cooperative behavior in physical and engineering systems has proven to be extremely valuable for understanding the organization and functioning of neural systems. At many levels of nervous system organization we observe complex spatio-temporal behavior related to the anatomical structure and the functional goals of the system. A substantial body of experimental evidence indicates that neurons often produce oscillations to achieve their functional goals, and in this manner act as dynamical systems and display their time asymptotic behavior in two ways: (1) they often produce statistically stationary time dependent activity whose mathematical image is that of an attractor, periodic or strange, and (2) they produce stimulus-dependent transient-dynamics which corresponds to trajectories in state space which are stable for a finite time.

"Attractor dynamics" is often found in the oscillations of small neural assemblies, such as Central Pattern Generators (CPGs), and this may also be found in some functions of the mammalian cortex. Transient behavior is observed in many sensory systems. Each of these styles of behavior are the result of synchronization and dynamical competition between different groups of living neurons. These phenomena and their varied manifestation at different levels of animal neural activity are the subject of this review.

Every activity of living animals is accomplished by particular neural assemblies. Rhythmic behavior such as swimming, running, breathing, and the like are controlled by small groups of neurons as members of a CPG. This class of neural assembly often utilizes as few as tens of neurons. Sensory systems range from a few thousand neurons in invertebrates to many millions in vertebrates. Yet another order of magnitude appears to be involved in mammalian learning and 'intelligent' activity. How can all these imperfect, nonidentical components, many of which exhibit chaotic oscillations when isolated, work together in a noisy, often unreliable environment? How do neurons produce specific stimulus-dependent or rhythmic behavior in such settings? To address these questions we must turn to the basic mechanisms of cooperative behavior among the neurons which comprise neural circuits.

The brain is a highly organized system in which its constituent units, the neurons, are connected in different fashions and hierarchical levels. The neurons in each subsystem usually have different morphologies, different intrinsic properties and, as a consequence, different activity patterns. We will discuss several mechanisms involved in the generation of rhythms and in the enterprise of information processing at three levels of the nervous system: i) in groups of small numbers of cells that control motor rhythms, the CPGs; ii) in ensembles of moderate number of neurons that perceive and relay external stimuli in olfactory systems; and iii) in very large arrays of neurons in the cerebellum.

2. Neural Oscillations

The membrane of neurons is a thin bilayer of lipids that isolates the extracellular medium from the intracellular medium. This barrier creates an electrochemical potential. Proteins inserted into the membrane form channels that control the inflow and outflow of ions. There are several active ion channels whose degree of permeability depends on the value of membrane electrical potential and on the concentration of particular ionic species such as calcium. The membrane potential of the neurons also changes as a func-

tion of the stimuli through the intrinsic biophysical properties of its passive and active ion channels.

The typical output of a neuron is an action potential or spike. This is a rapid change of membrane potential lasting about 1 ms. This high frequency electrical activity propagates along the body of the neuron and travels along axons to other neurons where it causes release of neurotransmitters that bind to the neuroreceptors of the receiving neuron in a chemical synapse. This causes a change in the local membrane potential of the receptor neuron, which in turn may develop its own action potentials and participate in the cooperative active of the connected neural assembly. In some cases the membrane voltage activity of a neuron can influence its neighbors by direct electrical connections or gap junctions of their membranes.

Several stimuli can arrive at the same time, and all of them are processed through the biophysical mechanisms of the membrane. When they are strong enough, another action potential is generated. The neuron's spiking activity can be sustained in time either in conjunction with the continued reception of the stimuli or, when the environment is appropriate, in its absence. Action potentials can be observed as isolated events, in repetitive firing activity with different frequencies and adaptation periods and on top of slow depolarizing waves of membrane potential called bursting activity. We display examples of such activity from our laboratory in figure 1. These time series come from intracellular recordings of the membrane potential of one of a pair of strongly electrically coupled PD or pyloric dila-

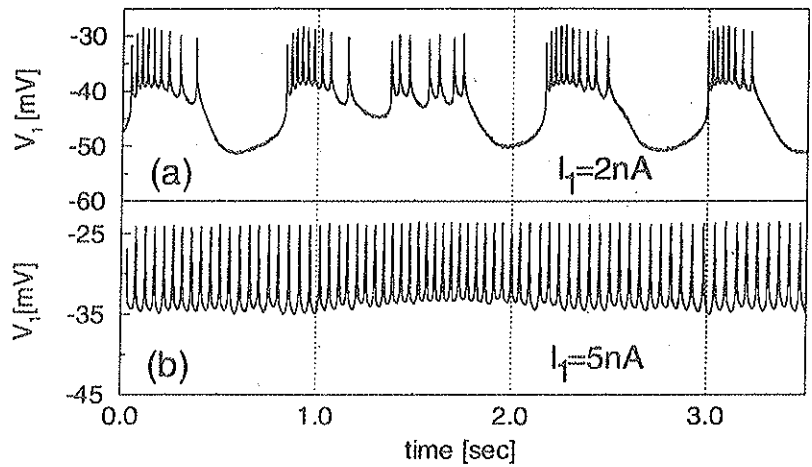


Fig. 1. Two different regimes of oscillations in a stomatogastric neuron obtained by applying two different injection currents. Top: bursting behavior. Bottom: spiking behavior (Elson *et al.*, 1998a).

tor neurons, from the pyloric CPG of the California spiny lobster, *Panulirus interruptus* (Elson *et al.*, 1998a). One of the PD neurons is deactivated by injection of a strong negative current. The other neuron is injected with a small positive external DC current I_1. This allows us to alter the behavior from chaotic bursting/spiking, figure 1a, to nearly periodic spiking, figure 1b, as well as to all intermediate oscillations.

This kind of spiking-bursting activity has been extensively modeled over the years. The membrane can be described as a set of isopotential compartments whose mathematical embodiment is obtained from the equivalent electric circuit of the portion of the membrane that they represent. We show this in figure 2. In these models the active channels are controlled by conductance variables described by the Hodgkin-Huxley (Hodgkin and Huxley, 1952) formalism. The detailed voltage dependence of the ingredients of these models can be obtained from experimental recordings in real neurons. Highly detailed models can be used to study the role of subcellular processes in generating the different firing patterns of individual cells and circuit activity. For large circuits, simplified integrate-and-fire dynamical models, substantially reduced from the details of the Hodgkin-Huxley descriptions can be used (Koch and Segev, 1998).

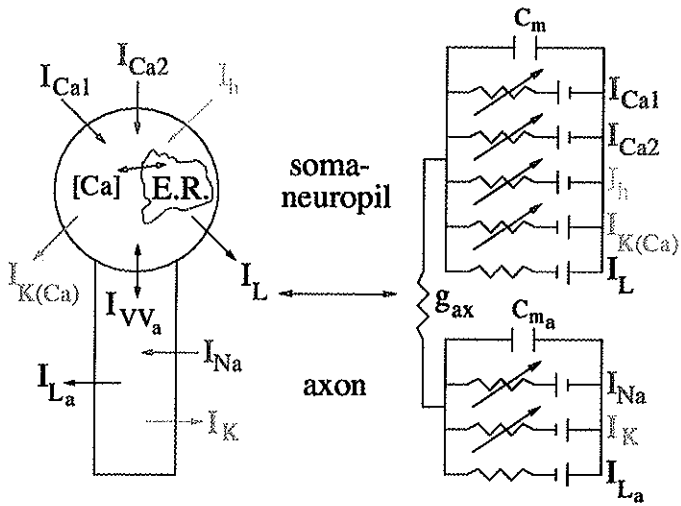


Fig. 2. A compartmental model of a stomatogastric neuron that includes a detailed characterization of the $[Ca^{2+}]$ storage and diffusion in the endoplasmic reticulum of the neuron. Capacitors represent the membrane capacitance, resistors represent ionic conductances (active and passive) and batteries represent reversal potentials for the different ionic species.

The typical spiking-bursting behavior of a model LP neuron from the pyloric CPG is shown in figure 7 of section 3.2. This model is used to study the role of the slow calcium dynamics in the genesis of the chaotic behavior and in the regularization mechanisms of the stomatogastric CPG neurons. This study requires the highly detailed model depicted in figure 2, with a two compartment architecture where the six active ionic currents used are distributed in the two compartments (soma-neuropil and axon) depending on their slow/fast evolution. A detailed calcium dynamics for the soma compartment includes Ca^{2+} storage in the endoplasmic reticulum and Ca^{2+} diffusion through the luminal and through the cytoplasmic membrane.

The integration of modeling and experimental techniques is useful to test hypothesis and draw important conclusions in the interpretation of physiological data from neural systems.

3. Synchronization and Competition in Minimal Circuits

In early seventies it was shown that the highly repetitive patterns of rhythmic motor activity of invertebrates could be sustained without any sensory stimulus, without any external influence or without neural signals from 'higher functions'. If a CPG is removed from an animal's body and placed in a salt solution that keeps the cells alive, this CPG may still generate essentially normal motor-output patterns for as long as many hours. There are several basic 'minimal circuits' of neurons that are known to generate characteristic oscillatory behaviors. This circuits are responsible for the fast onset of synchronous behavior, rhythmic activity and regularization of neural signals. For example, in central pattern generators, parallel inhibiting and electrical interconnections, and parallel inhibiting and exciting interconnections are encountered (Getting, 1989; Grillner *et al.*, 1991; Selverston *et al.*, 1997; Kristan, 1980). Thus to understand the origin of the synchronization phenomena in neural ensembles we have to begin from the simplest circuits: two neurons coupled together.

We discuss now several experimental results from our laboratory as well as the outcome from modeling this kind of essentially autonomous oscillation in individual or small collections of neurons taken from lobster CPGs depicted in figure 3.

3.1. *Experiments*

Experimental studies of synchronization in a pair of neurons that interact through naturally occurring electrical coupling have been reported in

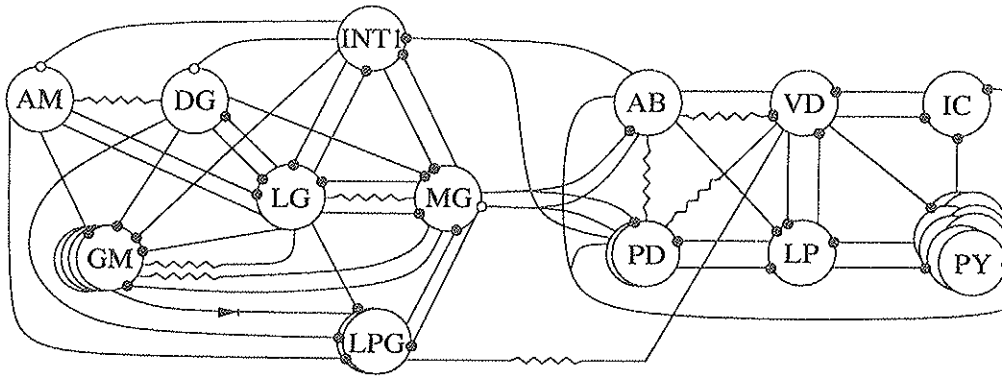


Fig. 3. Network organization in the pyloric and gastric CPGs in lobster. Types of neurons are indicated inside large circles (Selverston *et al.*, 1997). Solid circles indicate inhibitory connections and open circles excitatory connections. Connection from GM to LPG is rectifying electrotonic.

(Elson *et al.*, 1998a; Elson *et al.*, 1998b). Some of these experiments were carried out on two PD neurons from the pyloric CPG of the California spiny lobster. The strength of the natural electrical coupling can be altered during the observations of the preparation by use of a feedback device built for this purpose by N.F. Rulkov. The naturally occurring coupling characterized by a conductance of about 200 nS produces synchronization in the slow bursting oscillations, but the spikes are not synchronized by this coupling strength.

Individually, these neurons can generate highly irregular spiking-bursting activity figure 1a). Varying the control parameters injected DC current and interneuron conductance we found the following regimes of behavior (Elson *et al.*, 1998a; Elson *et al.*, 1998b).

Natural coupling produces state-dependent synchronization as shown in figure 4. With little or no applied current, the neurons fire spikes in irregular bursts in which the slow oscillations are well synchronized while the spikes are not; this is shown in figure 4a. Changing the magnitude and sign of electrical coupling restructures the cooperative dynamics. Increasing the strength of coupling produces complete synchronization of both irregular slow oscillations and fast spikes. Compensating the natural coupling of about 200 *nS* leads to the onset of independent irregular pulsations as in figure 4b. With net negative coupling, the neurons burst in antiphase, but now in a regularized pattern as in figure 4c. When depolarized by positive DC current, both neurons fire a continuous pattern of synchronized spikes

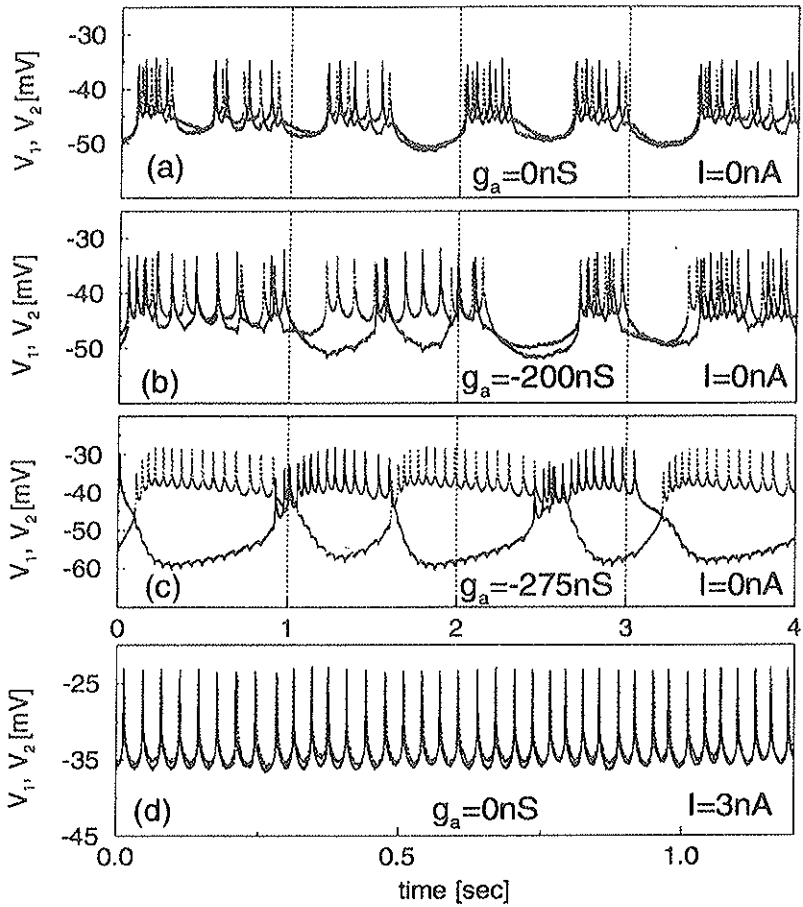


Fig. 4. Regimes of oscillations in two coupled PD neurons from the stomatogastric ganglion of the California spiny lobster for different coupling conductances g_a. The first three rows show the bursting behavior of the two neurons with different levels of synchrony. The last row shows synchronous spiking behavior (Elson *et al.*, 1998a; Elson *et al.*, 1998b).

as we show in figure 4d. In this figure g_a is the externally controlled conductance level.

The dynamics of slow oscillations changed as the external coupling conductance g_a was altered. With natural coupling $g_a = 0nS$ the slow oscillations stayed synchronized as seen in figure 5a even though each neuron displays very complex dynamics as shown in figure 5b or figure 4a. Additional dissipative coupling ($g_a < 0nS$) led to desynchronization. The desyn-

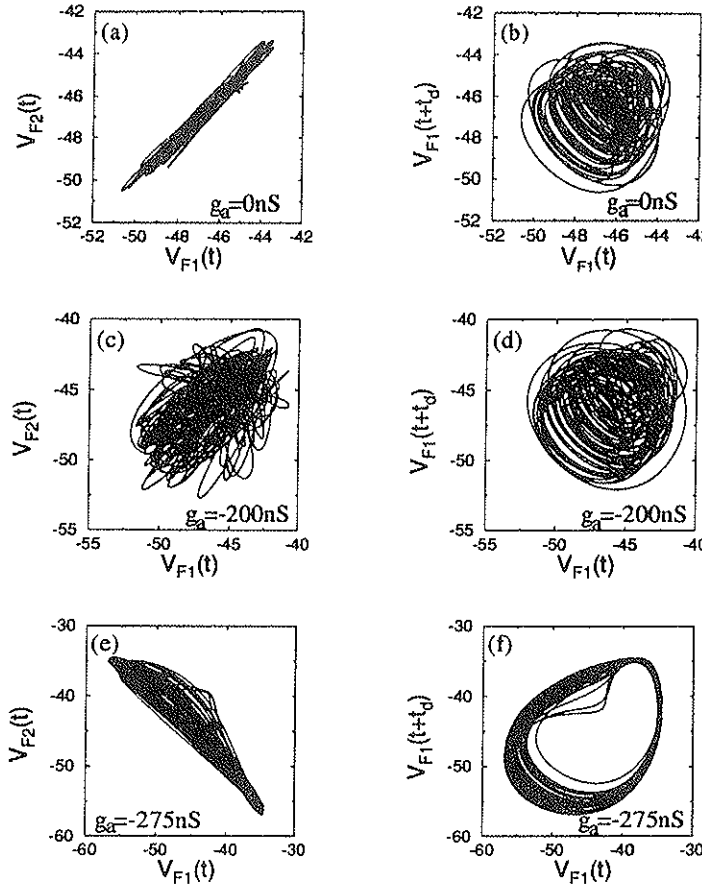


Fig. 5. Phase portraits of the slow components of oscillations in two coupled PD neurons as a function of the external conductance. The coordinates are the filtered membrane potentials of the two neurons (Elson *et al.*, 1998a) ($V_{F1}(t)$, $V_{F2}(t)$) in the left column. In the right column we display ($V_{F1}(t)$, $V_{F1}(t + t_d)$), with $t_d = 0.3$ sec. There is no externally injected DC current here.

chronized slow oscillations remained complex and aperiodic as we see in figures 5c and 5d; see also figure 4b.

Adding further negative coupling conductance which could represent an inhibitory synaptic connection caused the neurons to compete with each other and behave in an antiphase manner as seen in figure 5e. This regime of antiphase behavior was characterized by the onset of more regular, "almost periodic" bursts as we see in figure 5f.

We now consider the competition between neurons in more detail.

Spatio-temporal patterns of neural activity can be generated by competition mechanisms among the cells. Competition means that several units are active at the same time and through inhibition between the component neurons, even with simultaneous excitation, their states alternate as in the antiphase bursting of our two PD neurons coupled electrically with negative coupling as shown in figure 5e. The results from another experiment in our laboratory with two different neurons (LP and PD from the pyloric CPG with an inhibitory connection from PD to LP) show that the neurons burst in a nearly periodic alternating temporal pattern and their individual chaotic activity is regularized (Elson *et al.*, 1998a). When the polarity of one of the mutual connections is changed to excitation, regularization of the bursting behavior is lost.

Inhibitory synaptic connections between neurons appear to have a distinctive, even critical role to play in neuron assemblies. This type of connection between nonlinear oscillators is not typically found in physical systems, and this lesson from biology in itself represents an important new direction for the dynamical systems study of collections of nonlinear oscillators.

3.2. *Bifurcations and Modeling*

The experiments we have just described indicate that the slow bursting oscillations and the fast spiking oscillations of these two neurons have different thresholds for the onset of synchronization. This can be understood in terms of the different spatial sites of origin of the two types of voltage signal, the different mechanisms of synchronization, and the different conduction pathways and attenuation factors involved. The slow voltage oscillations that underlie bursting activity arise as a result of voltage-dependent ion channel activity in the membrane of neuropilar processes. The summed voltage signal will suffer some attenuation as it spreads by local current flow in the leaky cable array of the neuropil. However, two factors favor its effective transmission between the neurons: (a) the location of electrical coupling sites close to the site of slow wave generation, and (b) the slow time course of the voltage signal itself. In combination, these should allow a relatively strong and continuous interaction between the irregular slow oscillators. This mechanism resembles the synchronization seen in dissipatively coupled chaotic electrical circuits (Afraimovich *et al.*, 1986; Heagy *et al.*, 1994). In contrast, fast spike signals suffer strong attenuation as they spread between the spike initiation zone at the origin of the axon and the coupling sites in the neuropil. These factors argue for weak current flow between spike generators. If the spike generator of a neuron is close enough to its threshold, the transient current from the coupling pathway may drive

it to phase-locked firing. In electrical circuits, this type of chaotic pulse synchronization is known as threshold synchronization (Rulkov and Volkovskii, 1993). With natural coupling, these threshold mechanisms can synchronize spike activity in tonic firing but not in the bursting regime. When the neurons generate slow voltage oscillations, ion channels open in neuropilar processes, decreasing the membrane resistance. This shunts the spike-evoked currents as they flow in their coupling pathway, causing a failure in threshold synchronization.

As the strength of net coupling is decreased, the slow oscillations remain irregular with little change in waveform, but make a sharp transition from synchronous to asynchronous behavior (see figure 5). When the net coupling reaches an expected, negative conductance, the slow oscillations resynchronize in antiphase and become regular. These bifurcations argue for a dynamical origin of the irregular neuronal activity.

Based on these observations we have built a two compartment model of neurons from stomatogastric ganglion. The model incorporates six active ionic currents distributed in each of the soma-neuropil and the axon. It also accounts for *slow, intracellular* Ca^{++} *dynamics*. Two such model neurons, when electrically coupled, reproduce all five types of behavior found in our experiments and the transitions between the regimes are consistent with the observations (Varona, 1999).

The schematic description of our model of a PD neurons from the lobster pyloric CPG is shown in figure 6. We coupled two of these model neurons with an electrical, so-called gap junction coupling, to study how well our model neurons reproduce the results of experiments in our laboratory (Elson *et al.*, 1998a; Elson *et al.*, 1998b).

When the two model neurons are coupled with null or small coupling conductance, $g_{ec} \approx 0.001 \mu S$, independent chaotic behavior is observed. In (Falcke *et al.*, 1998) we present a detailed characterization of the chaos in the single neuron model as well as a detailed comparison of the model behavior with our experiments. Membrane potential bursts range from half a second to two seconds without periodicity as we can see in figure 7a. The number of spikes on the top of the slow bursting waves also changes from burst to burst. Note that local maxima of cytoplasmic calcium concentration ($[Ca^{2+}]$) mark the end of the burst plateaus. Calcium concentration inside the endoplasmic reticulum ($[Ca^{2+}]_{ER}$) evolves slowly, thus modulating in an anti-phase manner the faster oscillations of cytoplasmic $[Ca^{2+}]$ and influencing the length of the voltage plateaus. We will discuss the evolution of these three variables for different coupling strengths g_{ec}.

For all three cases discussed so far, small, medium and high positive coupling conductance, the bursting activity remains irregular regardless of

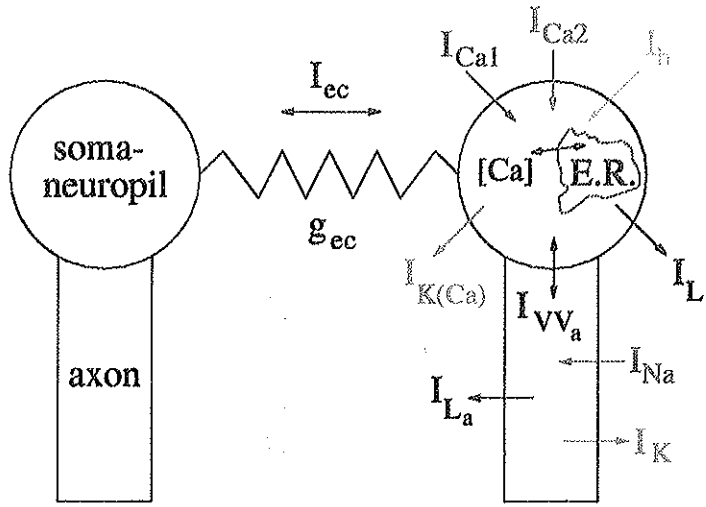


Fig. 6. Two stomatogastric model neurons that include a detailed characterization of the $[Ca^{2+}]$ storage and diffusion in the endoplasmic reticulum of each neuron. Each model neuron has two compartments: one represents the soma-neuropil activity, and the other represents the axon including the spike generating zone of the neuron. Except for the critical addition of the Ca^{2+} dynamics, this is a simple extension of many Hodgkin-Huxley models of this class of neuron. If this calcium dynamics is absent or the concentration $[Ca^{2+}]$ is fixed at some value, the model neuron does not exhibit chaotic oscillations, and it equally does not reproduce the behavior of these neurons.

the degree of synchronization. Thus, synchronization occurs without regularization. When the two neurons are coupled with a small negative conductance $g_{ec} = -0.001\mu S$, thus inverting the sign of the current coming from the electrical coupling in both neurons, anti-phase synchronization is observed in the membrane potentials as seen in figure 7d. Furthermore, the two neurons regulate their bursting behavior in the sense that the lengths of the burst are kept uniform. Note in figure 7d that $[Ca^{2+}]_{ER}$ remains nearly constant for the two neurons, while $[Ca^{2+}]$ oscillates regularly but in anti-phase with respect to the other neuron. In the previous cases $[Ca^{2+}]_{ER}$ oscillated slowly with a large amplitude. In our model, chaotic behavior is sustained in the single neuron model whenever $[Ca^{2+}]_{ER}$ oscillations are present. If $[Ca^{2+}]_{ER}$ is kept constant, the model produces regular bursting activity. For a small negative electrical coupling, the calcium dynamics in the ER of each neuron is maintained constant, since the fast oscillations of calcium in the cytoplasm are rapid enough and regular enough to have no influence on the slower calcium diffusion through the endoplasmic reticulum membrane. Again, if the calcium concentration in the ER is kept constant, regu-

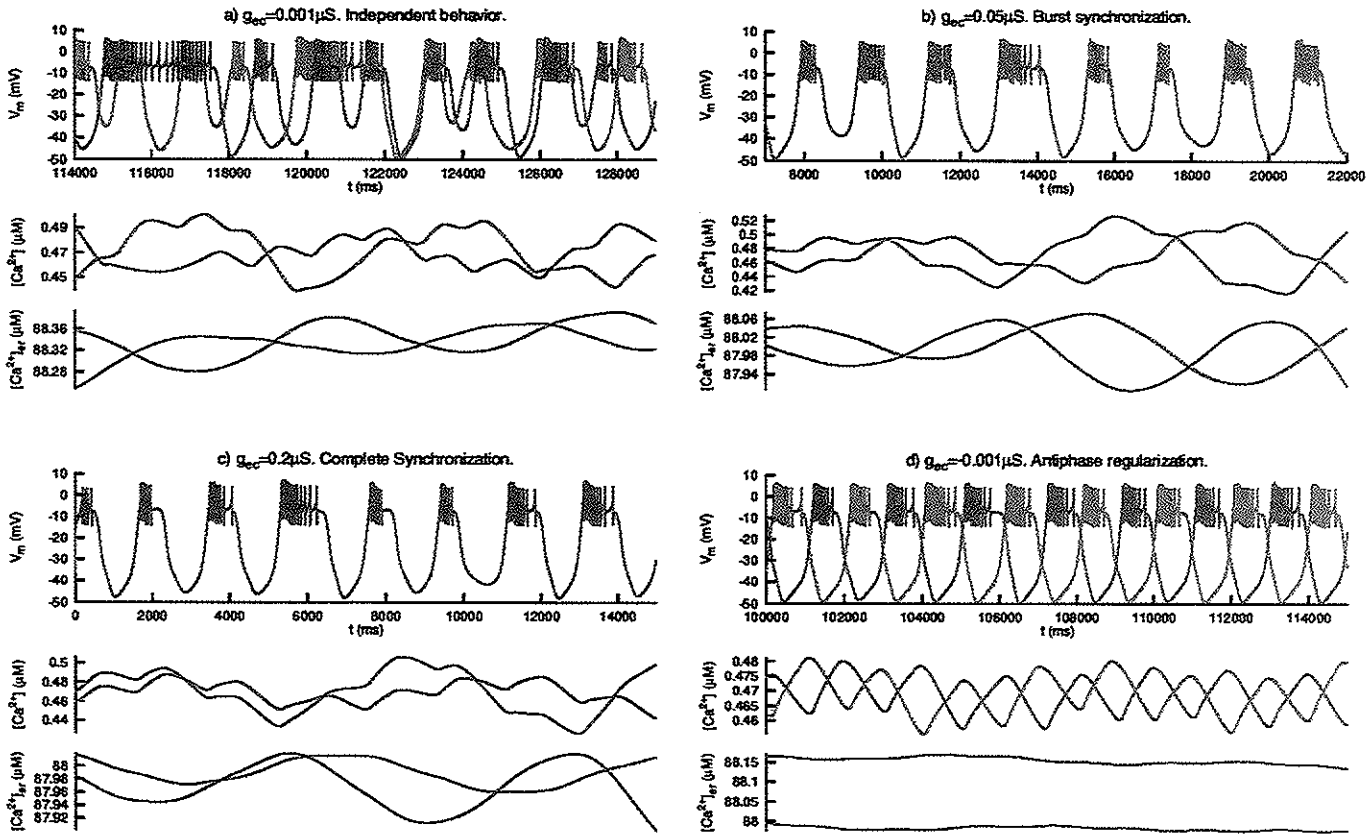


Fig. 7. Four different collective behaviors observed when two PD model neurons are coupled electrically as in figure 2 of (Elson *et al.*, 1998a): a) independent chaotic bursting activity arising when $g_{ec} = 0.001\mu S$, b) burst synchronization associated with $g_{ec} = 0.05\mu S$, c) total synchronization which appears when $g_{ec} = 0.2\mu S$, and d) anti-phase synchronization with regularization which comes when $g_{ec} = -0.001\mu S$. Activity for neuron one is plotted with a dark trace, while neuron two is represented with a light trace. In each of the graphs we display, from top to bottom: membrane potential V_m, cytoplasmic calcium concentration $[Ca^{2+}]$, and calcium concentration inside the endoplasmic reticulum $[Ca^{2+}]_{ER}$. This model

larization of the chaotic behavior occurs. This behavior is also observed when the regularization is obtained by periodic driving through small periodic pulses of current injection and when the two neurons are coupled with mutual inhibitory chemical synapses.

3.3. *Lessons from Modeling Small Assemblies*

In this section we described both our work on individual neurons based on experiments using the pyloric CPG from the stomatogastric ganglion of the California spiny lobster and our modeling and experiments on small subcircuits of this CPG. We found that the *inhibitory* synaptic couplings appearing in the natural networks were essential for producing the regulation of the chaotic oscillations prevalent in the dynamics of the isolated neurons. Further we identified the need for an additional slow dynamical process, beyond the traditional Hodgkin-Huxley ionic currents if we wish to account for the chaos observed in individual neurons. In our modeling we suggested that this additional slow dynamics is due to calcium exchange between the intracellular medium and the endoplasmic reticulum. This is now being tested in various laboratories, including our own, where we are trying to establish the qualitative presence of this kind of calcium dynamics.

Starting with the description of an individual neuron, we established further arguments for the kind of modeling we have performed by comparing the observations on two electrically coupled PD neurons from the pyloric CPG to our modeling of the same situation. In the experiment as well as in the model we established a clear quantitative comparison between the two. It is important to note that the experiments also included variations of an externally imposed conductivity so the range of dynamical behavior described in the model and seen in the experiments was quite broad: from synchronized spiking firing to out-of-phase oscillation to unsynchronized chaotic oscillations.

To this point then we have established a clear set of models whose biological ingredients are clear and testable and whose dynamical behavior we understand. Working with this knowledge we are now going to move on to explore the importance of inhibitory connections among neurons in a wider arena. We shall find, as indicated in our introduction, that both the details of the neural dynamics of each component and the architecture associated especially with inhibitory connections in neural assemblies provides a suggestive basis on which we may understand complex and rich behavior of these assemblies, and further, the model behavior, at least in a qualitative sense reflects phenomena known from many observations on such assemblies.

4. Sensory Dependent Dynamics of Neural Ensembles

4.1. *Microcircuits with Antagonistic Coupling*

When one of our small circuits viewed as a dynamical system has more than one attractor for the same settings of its biological parameters, we have a situation of multistability. Which attractor 'wins' depends on the initial conditions of the system, and we can 'reset' those initial conditions by applying stimuli of various characteristics. We discuss here the ability of oscillating neural circuits to switch between different states of oscillation in two basic examples. We model two quite distinct small neural circuits, which are presented in figure 8.

Figure 8 (left) shows a neural couple from the lobster stomatogastric ganglion (STG) (cf. LG-MG connection in figure 3) (Selverston and Moulins, 1987), and figure 8 (right) a typical vertebrate thalamocortical (RE-TC) circuit (Steriade *et al.*, 1993). Although the functional role played by these circuits is very different, the presence of antagonistic couplings between different parts of the circuit makes them exhibit common dynamical features. The STG circuit is composed of two neurons coupled via both gap junction and inhibitory synapses. The second consists of coupled pairs of interconnected thalamocortical relay and thalamic networks passing sensory information to the cerebral cortex. Both circuits have contradictory coupling between symmetric parts. The thalamocortical model has excitatory and inhibitory connections and the STG model has reciprocal inhibitory and electrical coupling. We describe the dynamics of the individ-

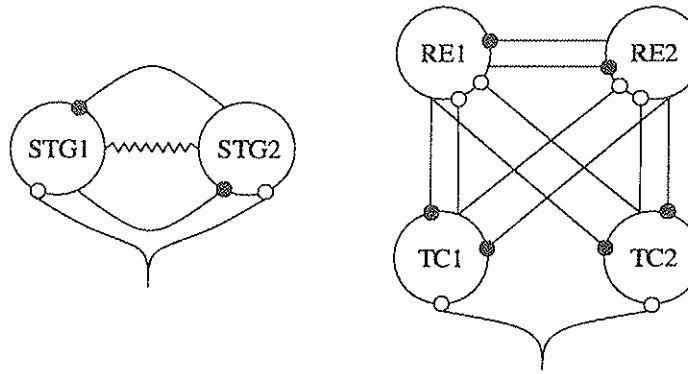


Fig. 8. Two basic neural microcircuits. Left, a STG circuit and right, a thalamocortial circuit. Solid circles indicate inhibitory connections and open circles excitatory connections. The resistor symbol in the STG circuit denotes a gap junction, an electrical connection between the two neurons.

ual neurons in these circuits by conductance based ordinary differential equations of Hodgkin-Huxley type. Both model circuits exhibit bistability and hysteresis in a wide region of coupling strengths. The two main modes of behavior are in-phase and out-of-phase oscillations of the symmetric parts of the network.

We investigated the response of these circuits to trains of excitatory spikes with varying interspike intervals T_i and with quite small amplitude pulses. These are a simple representation of spike trains received by the basic circuits from sensory neurons. Circuits operating in a bistable region are sensitive to the T_i of these excitatory inputs. T_i variations lead to changes from in-phase to out-of-phase coordination or vice versa. The signaling information contained in a spike train driving the network can place the circuit into one or another state depending on the interspike interval. It is important to note that this happens within a few spikes, and then these states are maintained by the basic circuit after the input signal is completed, and the circuits remain in the reset state until a new spike train of another T_i is received. When a new signal of the correct T_i enters the circuit, the state can switch again. Our main results are presented in figure 9. See (Rabinovich *et al.*, 1998a) for further details.

Bistability occurs when there are two distinct solutions to the conductance based differential equations describing the circuit that coexist over a range of settings of the various parameters in the equations. In (Rabinovich *et al.*, 1998a) we explored a range of electrical couplings over which the STG circuit had two distinct solutions and we investigated a range of the strength of the inhibitory coupling over which the RE-TC cells act in the same fashion. In the state space of the systems we see two distinct orbits or phase portraits for the two solution sets. These represent two distinct attractors for the dissipative neural dynamics. Whether, after an initial transient behavior, the circuit ends up on one attractor or another depends on the initial conditions for the solution of the differential equations. In state space each attractor has a set of initial conditions that bring the solution to it, and this collection of initial conditions is called its basin of attraction. Figure 10 shows the two attractors for the RE-TC system in the same state space. As one can see, the two attractors are quite close in this space, supporting the fact that transitions between them can be easily induced by the periodic spike trains we introduce.

Two potential uses may be made of the reset capability of bistable circuits. First, in lobster STG circuits it is known that neuromodulators can alter the character of neural oscillations in accordance with selected functional behavior (Selverston and Moulins, 1987). The reset capability of sensory spike trains may also be used to achieve this goal. Second, this reset

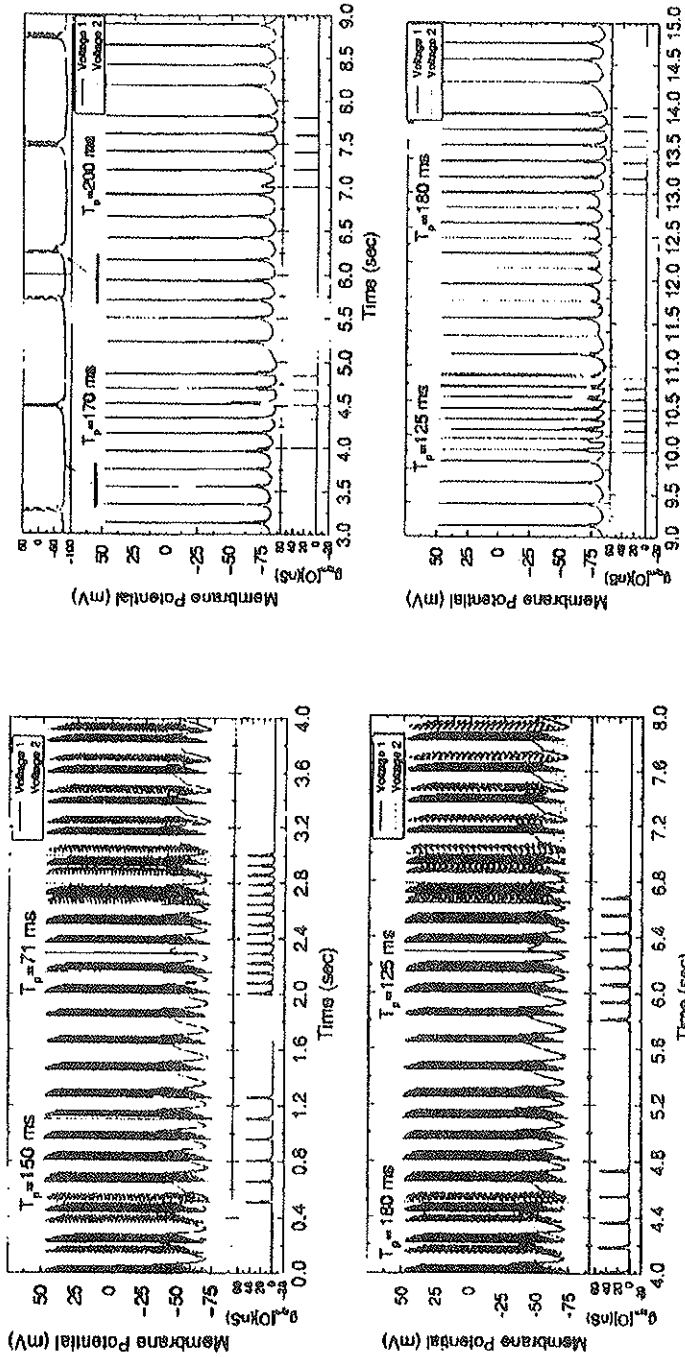


Fig. 9. Left: time series showing the effect of 1*s* of the periodic external forcing in the CPG circuit at several values of T_p: 150, 71, 180 and 125ms. Right: the effect of 1*s* of the periodic forcing in the RE-TC circuit at the same values of T_p as in the CPG circuit. There is an expanded time scale view of the membrane voltage in the RE-TC circuit above the upper panel.

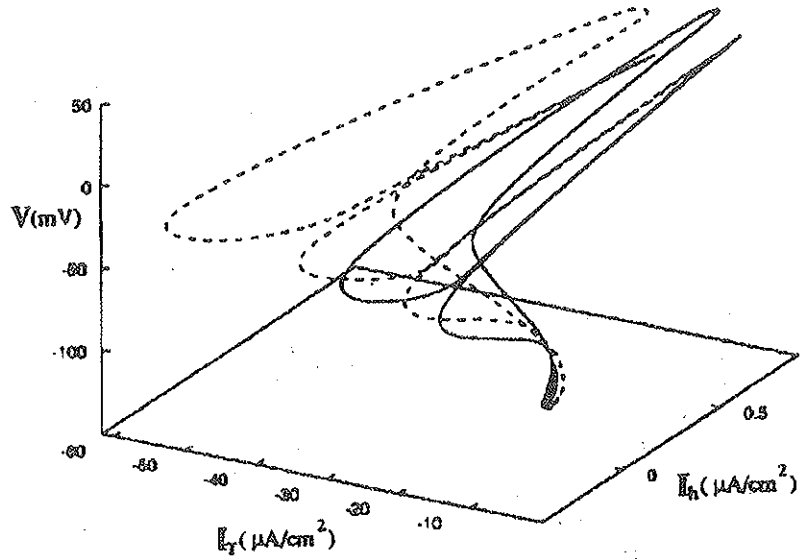


Fig. 10. State space portrait of the two coexisting attractors for the RE-TC system. The solid line is the orbit in $[V(t), I_T(t), I_h(t)]$ space of the in-phase oscillations. The dotted line is the path taken in the same state space by the out-of-phase oscillations. The closeness of the two attractors leads to the ease with which spike trains with appropriate T_p can induce transitions between them.

capability may be a way in which neurons interpret information coming from sensory sources and reformat it for use further along in the animal's processing and decision system. If this "learning" function is correct, the mechanism could potentially be useful in short term memory where more complex circuitry would be reset for such a purpose.

The behavior of these two basic neural circuits are examples of what we call "calculation with attractors". In the phase space of a neural assembly the number of such attractors can be very large, and if their basins are widely enough separated such a system could function as an associative memory (see Rolls and Treves, 1998 for a review and references). Recent experimental results show that "calculation with attractors" is not so typical for sensory systems. Let us discuss this problem in more detail using the olfactory system as an example.

4.2. *Transient Dynamics in Olfactory Systems*

Consider the first stages in odor processing of the olfactory systems, namely the actions of the olfactory bulb (OB) in vertebrates or antennal

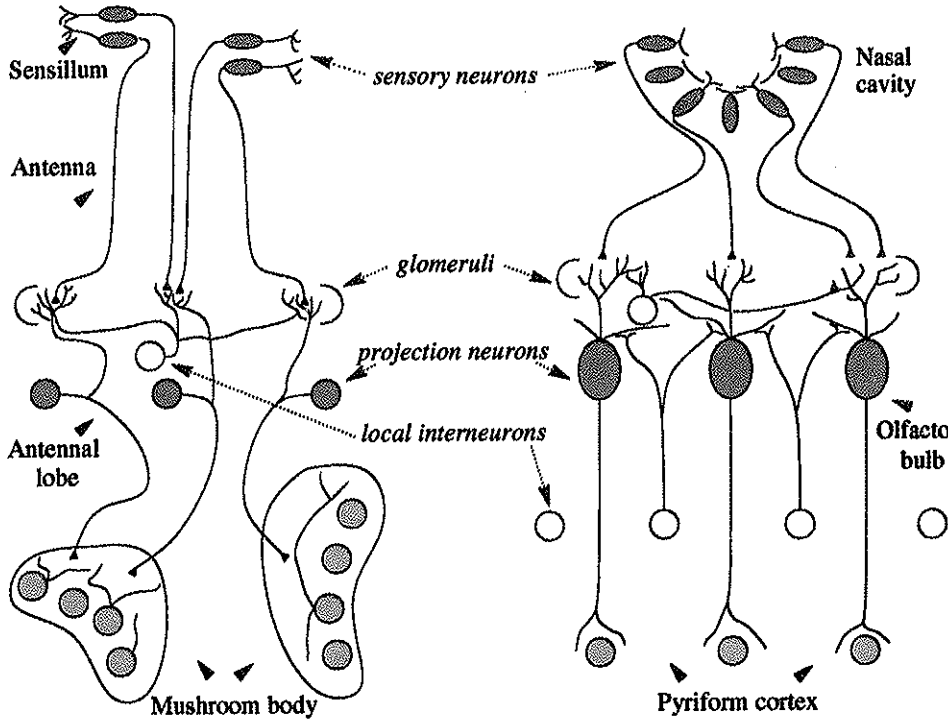


Fig. 11. The principal neural circuits in the olfactory systems of insects (left) and vertebrates (right).

lobe (AL) in insects. The OB or the AL receives information about odors from the sensory neurons through special terminals or junction boxes called glomeruli, and then they reorganize the spatial information associated with the sensory spike trains in spatiotemporal way and present it to the cortex in vertebrates or the mushroom body in insects; we depict this in figure 11. The OB or AL plays the role of a "contraster" and "amplifier" of the sensory information. Both OB and AL consist of two types of neurons. The first are representative or projector neurons which send the information to the cortex. These are PN neurons in the AL and mitral and tufted cells in the OB. The other type is an interneuron (IN) which acts within the OB or AL itself. The PN are excitatory neurons, and the IN are inhibitory.

The complex inhibitory connections in OB or AL are extremely important for the organizing the spatio-temporal representation of the sensory stimuli. In particular, this connection establishes complex antagonistic interactions between PN cells with the goal of forming an "odor oppo-

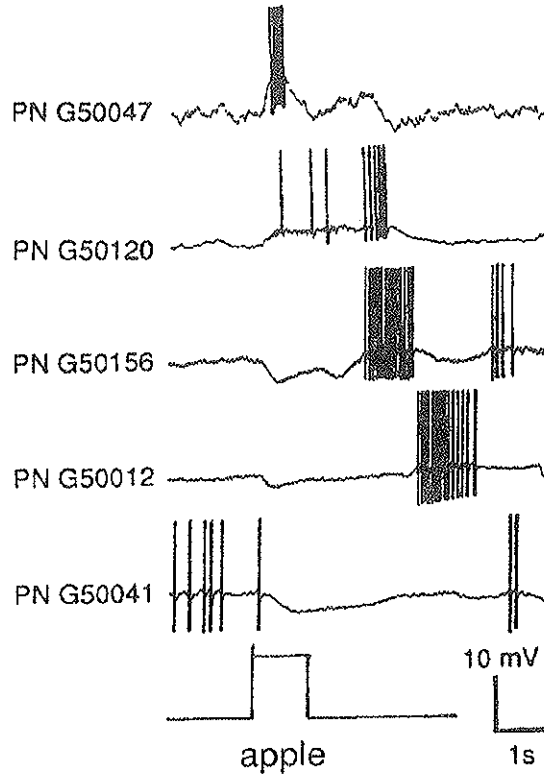


Fig. 12. Range of temporal patterns of response to a single odor across neurons. Temporal response patterns of five different antennal lobe PNs in response to the odor of apple. The recordings (all intracellular) were performed sequentially in the same animal over a 3.5 hr period. Traces have been aligned on the odor pulse. From (Laurent *et al.*, 1996).

nency" mechanism, thus preventing certain odor conditions being reported simultaneously (Pearce, 1997). In nonlinear dynamics language this "odor opponency" arises from competition between the PN cells. Such mechanisms could be similar to the "color opponency" mechanism found in color perception. In the visual system, ganglion cells may be excited by one particular frequency, yet inhibited by another (Mori and Shepherd, 1994).

Competition between groups of PN cells is important not only for the simple differentiation of distinct odors, but also for the representation of a single odor in a spatio-temporal fashion. As experiments (Laurent *et al.*, 1996) and modeling (Rabinovich *et al.*, 1998b) have shown, AL represents incoming sensory information to the mushroom body using both spatial and

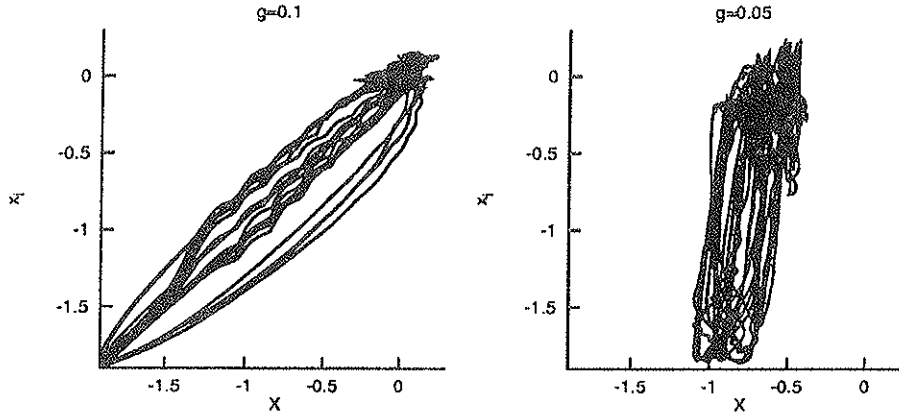


Fig. 13. Activity of a single HR unit x_i *versus* the average activity X ($X = \frac{1}{M}\sum_{i=1}^{M} x_i(t)$, where M is the number of neurons in the cluster) for two values of the coupling g.

temporal competition as well as temporal synchronization between PNs. Such odor-specific temporal behavior is enormously robust against noise and remarkably reproducible when the initial conditions of the circuit are different when the stimulus arrives. An example of the range of temporal patterns of response to a single odor across neurons is shown in figure 4.2. We hypothesize the inner inhibition of AL is responsible for the temporal encoding of the sensory information in a "winner-less competition" fashion.

4.3. *Lessons from Sensory Dependent Neural Dyamics*

Simple models of basic neural circuits taken from widely different areas of use show substantially the same bistability in their behavior. In coming spike trains can select one out of many possible attractors of such a system by varying their interspike intervals which effectively resets the circuit in the basin of attraction of one attractor or another. Short term 'memory' can be accomplished by setting a circuit in some attractor and leaving it there as long as required to achieve some desired behavior.

Competition among neurons in an assembly through an architecture of inhibitory connections allows robust identification of incoming sensory signals as well as allowing them to be formatted for further processing by higher cortex functions. Reliable signal identification occurs for a very large range of initial conditions and appears quite robust against external as well as internal noise in the neural circuitry.

5. Synchronization of the Mean Field in Large Neural Assemblies

5.1. *"Coarse Grained" Dynamics of Chaotic Neurons*

The emergence of spatio-temporal order or coherent structures in heterogeneous neural ensembles having individual elements with irregular behavior is one of the most intriguing problems in neuroscience. Over the last decade spiral patterns in the cortex of animals and humans have been observed (Prechtl *et al.*, 1997; Shevelev *et al.*, 1992). We have investigated this phenomenon of 'order creation' using a simple, well tested model of a two-time-scale chaotic neuron.

We consider a lattice of N different chaotic neurons electrically coupled to their near-est neighbors. We select as the component element in the network the three dimensional Hindmarsh-Rose (HR) (Hindmarsh and Rose, 1984; Wang, 1993; Huerta *et al.*, 1997) model neuron. We describe a two-dimensional lattice composed of such HR elements by the equations

$$\frac{dx_i}{dt} = y_i + ax_i^2 - x_i^3 - z_i + e_i - g\sum_j (x_i - x_j) \tag{1}$$

$$\frac{dy_i}{dt} = b - cx_i^2 - y_i \tag{2}$$

$$\frac{1}{\mu}\frac{dz_i}{dt} = -z_i + s(x_i + d). \tag{3}$$

The index i runs over [0,N], and the index j runs over the four nearest neighbors of unit i. The constants a, b, c, d, s, e_i, and μ are model parameters in which $\mu \ll 1$ gives rise to slow bursting dynamics, and g is the homogeneous coupling strength among neighboring units. Numerical simulations of two-dimensional lattices built up with heterogeneous HR neurons show that cooperative behavior among the elements produces large-scale coherent structures with slow periodic oscillations, even though the circuit is built from different neurons which are individually chaotic. We show such a coherent pattern in figure 14. The parameter e_i is chosen from a random collection to make the individual components of the network behave differently.

In order to understand the origin of these large-scale coherent structures we investigated the cooperative behavior of a cluster of such chaotic neurons (Rabinovich *et al.*, 1999). We found a striking new phenomenon: when the size of this cluster is sufficiently large the average activity is regu-

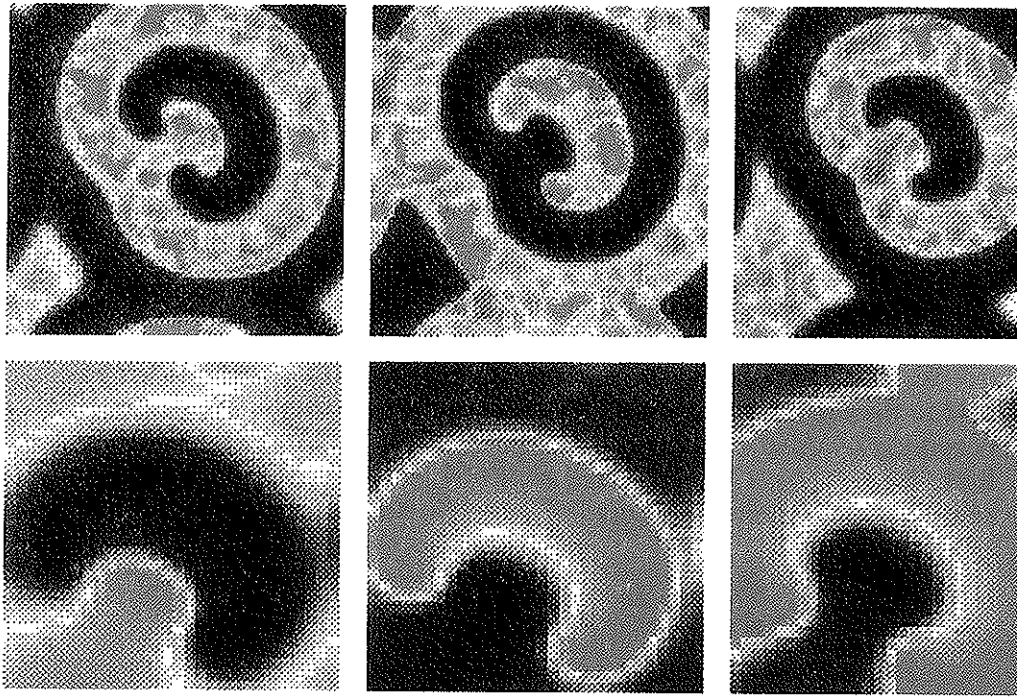


Fig. 14. Top row: evolution of a periodic spatio-temporal pattern observed in a network of 100x100 Hindmarsh-Rose elements. Bottom row: periodic spatio-temporal patterns observed in a network of 30x30 coarse grained elements.

larized. In contrast, small groups of neurons clearly exhibit three different kinds of chaotic dynamics depending on the value of the diffusive coupling g: (i) well developed chaos whose dimension increases with the number of chaotic neurons for a small value of the coupling, (ii) chaotic synchronization of the burst oscillation for moderate coupling, and (iii) complete chaotic synchronization of both spikes and bursts for strong coupling (Afraimovich *et al.*, 1986; Pecora and Carrol, 1990).

The dynamical mechanism leading to ordered average behavior of the cluster relies on synchronization and regularization of the activity of the neurons inside the grain. The degree of synchronization of a single neuron with the average activity of the whole cluster depends on the strength of the coupling, as one can see the left panel of figure 13. In the case of regular behavior, when $g \approx 0.1$, single neuron activity is highly synchronized with the periodic mean field. For $g \approx 0.05$, the synchronization between mean field and individual behavior disappears, and one observes spatio-temporal disor-

der as is visible in the right panel of figure 13. Thus, for a moderate value of g the cluster of neurons behaves as a single element with periodic slow dynamics. This mechanism for the regularization of oscillations in disordered neural assemblies could be a quite general principle for many systems.

5.2. *Regular Spatio-temporal Patterns in Disordered Neural Assemblies*

Now we can explain the existence of regular spatio-temporal patterns in neural ensembles which extend over a very large number of individual component neurons. First, the existence of such large-scale structures is not possible in weakly diffusive ensembles because the local oscillations of neighboring neurons are not correlated for small couplings g and the mean field of the assembly becomes homogeneous and stable. For moderate values of the coupling the coarse grain assembly should exhibit regular spatio-temporal patterns. As confirmation of this conjecture, we have checked the behavior of a network consisting of coarse grained units with slow periodic behavior. The results are shown in figure 14.

Thus, the formation of large-scale coherent structures in neural networks consisting of HR chaotic neurons with fast and slow oscillations exhibits two key features. The first is the regularization phenomena inside clusters of chaotic neurons, *i.e.*, the coarse grains. This regularization of the behavior results from the action of the average activity of fast pulsations in the slow coarse grained dynamics. The second feature is the instability of the homogeneous oscillation modes in a neural network considered to be a collection of coarse grained elements.

6. Synchronization and Competition in the Activity of the Inferior Olive

We begin with a few words about the role of dynamical modeling for understanding neural activity. When we talk about different neural systems, modeling may have quite different goals. For example, it is clear what kind of job is required of the antennal lobe. It must contrast and amplify the information about the odor as presented to it by the sensory neurons, and then it must send this information to the mushroom body for further processing to accomplish recognition and storage. We do not presently have detailed anatomical knowledge about the architecture of the antennal lobe but we can still model this clear function, and we can establish several general ways to realize this function. The situation is not always so. For instance, the architecture of the inferior olive and the cerebellar circuits of mammals has been investigated anatomically and physiologically in great detail. How-

ever, the functional role of these complicated circuits is still unclear (Llinas *et al.*, 1997). In contrast to the situation in olfactory processing, our challenge is to utilize this detailed knowledge and understand what such a system is able to do. Now we try to address this answer this question.

The inferior olive (IO) has been proposed as a system that controls and coordinates different rhythms through the intrinsic oscillatory properties of the individual IO neurons and the modulated inhibition by another set of neurons in this system: the cerebellar nuclei (Llinas and Welsh, 1993). We show this in figure 15. The inferior olive cells are electrically coupled and have strong oscillatory activity. Their axons transmit synchronous and rhythmic excitatory synaptic input to both the cerebellar nuclear cells and to the Purkinje cells of the cerebellar cortex. The phasic response of the Purkinje cells is transmitted as inhibitory inputs to the cerebellar nuclear cells. Thus the nuclear cells are excited by the inferior olive cells and later inhibited from the Purkinje cells. This inhibition leads to rebound excitation. The nuclear cells also send an inhibitory feedback to the inferior olive, thus closing this loop.

The IO neurons generate subthreshold oscillations and spiking activity as shown in figure 16. The time series shown in this figure come from the model IO neurons used in constructing a network of such cells connected with gap junctions among close neighbors. These networks incorporate a simple inhibitory feedback that implements the action of the cerebellar nuclei neurons. The model is realistic enough to generate subthreshold oscillations as well as spiking behavior in the amplitude and frequency ranges reported for the inferior olive neurons. Different oscillation frequencies can be obtained by applying a constant DC current to the model neurons. Network architectures were built up to 200x200 neurons connected electrically to their nearest neighbors and with an inhibitory feedback from the nuclei. These inhibitory connections are modeled without neither the detailed implementation of the other cell types involved in the inhibitory loop (Purkinje cells or cerebellar nuclei). A simple integrate and fire unit takes into account whether a group of neighbor IO neurons (typically three to nine) have a synchronous spiking event (in a time window of 5ms) and it evokes a delayed IPSP in this small cluster of IO neurons. Then the integrate and fire neuron has a refractory period where it can not fire for a short time. The real circuit and our simplification of the inhibitory loop are shown in figure 15. With this architecture, the network is able to generate spatio-temporal patterns as those shown in figure 6. Sequence goes first from left to right in line 1 and continues in line 2. Regions with the same color have synchronous behavior. Light colors mean depolarized potential. The spatio-temporal patterns consist of propagating wave fronts of spiking activity that can remain bounded in a region of the network.

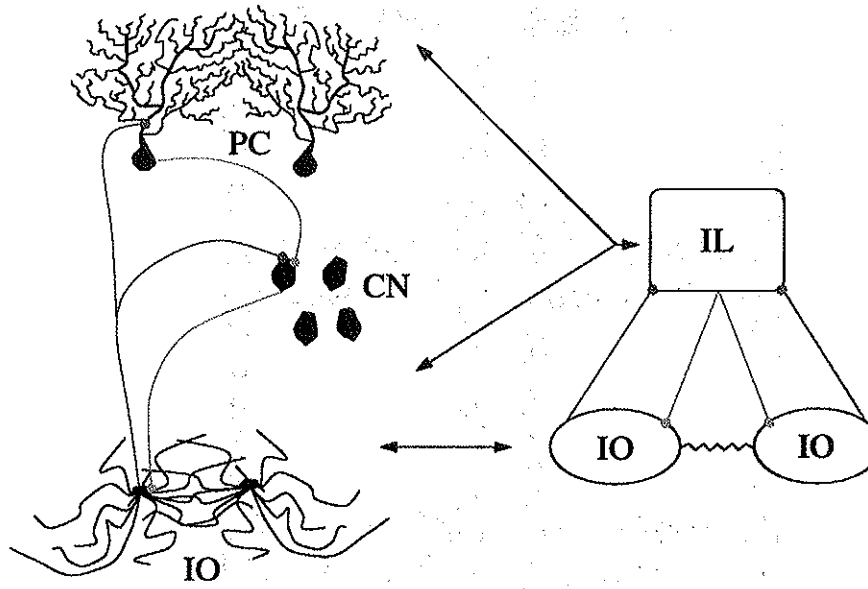


Fig. 15. Left: Representation of the cerebellar inhibitory loop. IO: inferior olive neuron; CN: cerebellar nuclei; PC: Purkinje cell. Dark connections are excitatory, light ones are inhibitory. Right: schematic simplification of the inhibitory loop (IL) used in our model.

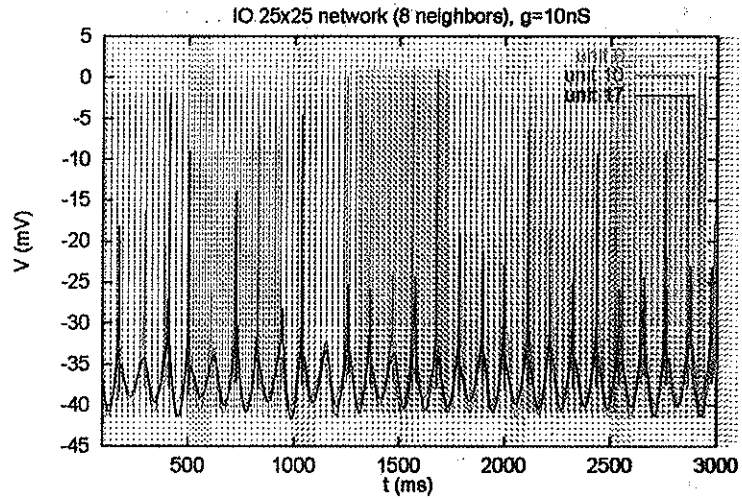


Fig. 16. Time series of the subthreshold and spiking activity in a model of the inferior olive. The membrane potential of three different neurons members of an ensemble electrically coupled is shown. There is a high degree of synchronization caused by the gap junctions.

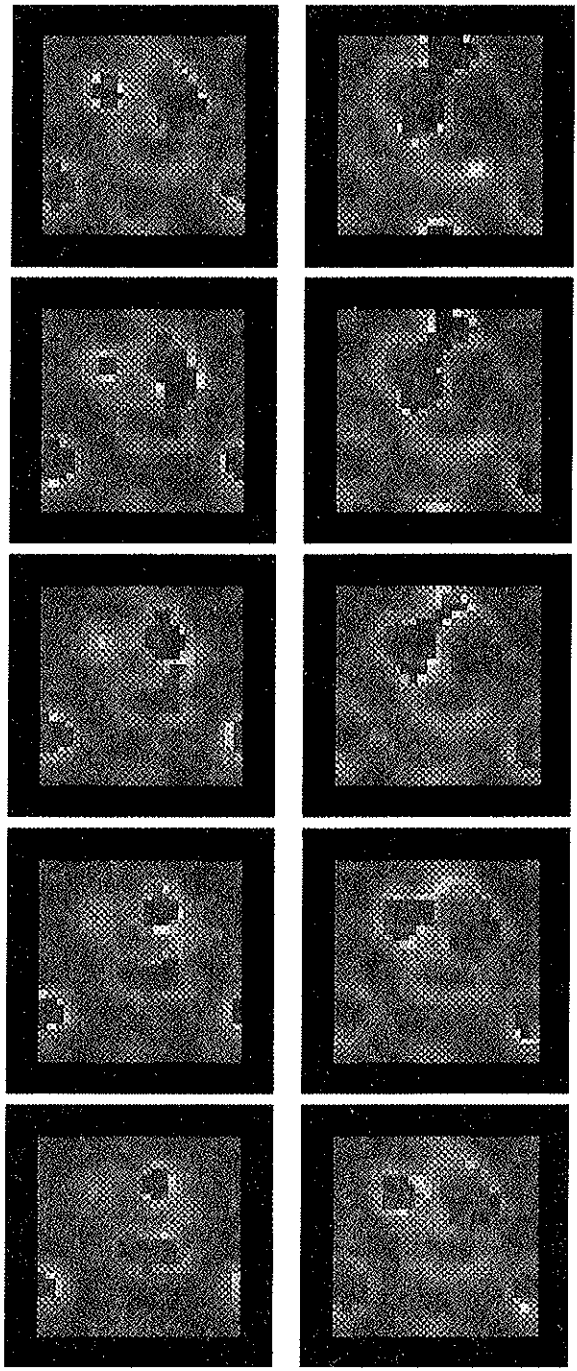


Fig. 17. Spatio-temporal patterns generated with a network of 30x30 inferior olive neurons electrically coupled to their nearest neighbors. Sequence goes from left to right and from top to bottom. Regions with the same color haver synchronous behavior. Light colors mean depolarized potential.

We have investigated several ingredients that modulate the frequency of the spiking behavior in the network. Three major factors have been identified:

— The electrical coupling conductance; higher values of the coupling strength increase the synchronization level and diminish the frequency of the spiking behavior.

— The number of electrically coupled neighbors also decreases the frequency of the spiking behavior, but only for a strong enough coupling. In this case the degree of synchrony among cells is higher although the frequency of the subthreshold oscillations remains constant under all these changes.

— The presence of inhibitory chemical synapses coming from the cerebellar nuclei changes both the spiking frequency and the frequency of the subthreshold oscillations. For a given value of the coupling conductance and a fixed number of nearest neighbors, the inhibition usually decreases the frequency of the spiking activity. Some simulations show that the frequency can also increase depending on the time constant used to implement the inhibitory synapses.

We use this model to understand how the IO oscillations can encode and control several simultaneous rhythms. The intrinsic modulation of the input on the system activity, the importance of the coexistence of the synchronization induced by the electrical coupling, and the competition caused by the inhibitory loop have been identified out by the use of this model. Realistic models such as this can also explain the role of some of the subcellular processes and the particular neural physiology in the generation of observed rhythms.

7. Discussion

We have illustrated how neural circuits on different levels use synchronization and competition mechanisms for the production of rhythms and for information processing. Several important questions come to mind when we want to understand how general the these mechanisms can be.

— What morphological and physiological characteristics of individual neurons and their interconnections are essential to perform a specific cooperative function?

— What is the significance for the generation of a rhythm or for information processing of the dynamical heterogeneity of neurons?

— Do different circuit architectures underlie different processing functions?

We briefly discuss these questions.

In the last few years many neuroscientists have agreed that "the traditional views of the significance of single neurons are fading in power" and "there are signs from experimental and theoretical work on the neocortex that we are on the threshold of a revolution in which the hegemony of the single neuron will be replaced by much more circuit-oriented concepts" (Douglas and Martin, 1991).

Do we agree with this viewpoint? Well, yes and no! "Yes", because in fact many experiments show that very often groups of neurons organized in microcircuits behave as a dynamical unit. However, neurons are individually complex, capacitive and nonlinear devices that transform streams of neurochemical packets into electrical waveforms. Their modes of operation are intrinsically time dependent, and, therefore, their functions or role in a circuit cannot be limited by their position in the circuit architecture. In particular, there are many examples that suggest the importance of the details of individual neural dynamics: in stimulus representation the neural activity of learning and storage requires neurons with specific features such as subthreshold oscillations; coding of the time delay between several inputs, etc. However, in general, the microcircuit concept is very useful and powerful (Rolls and Treves, 1998).

Of course, all neurons are different, but as we showed above both the synchronization and competition phenomena are structurally stable. This observation implies that variation of individual neuron parameters inside the group that produces cooperative behavior can be large, yet the functional requirements of the assembly are well met as a result of the architecture. Real neural systems may achieve some of this stability as the result of some redundancy as well.

Does a neural circuit have to possess a unique architecture to achieve a specific function? If so, strict rules are presently unknown. Circuits with different architectures are able to do the same job (Getting, 1989; Ullstrom *et al.*, 1998). Conversely, circuits with fixed structure may play different roles when receiving different inputs. Even for the same input arriving at different times, the outcome may differ as the responses of neural circuits may be quite state dependent.

Our view of neural assemblies as input/output dynamical systems has led us to uncover some quite general principles of neural activity. In the examples presented in this paper, these include the essential role of inhibitory connections among individual, often complex, nonlinear neural oscillators. In our discussion much of the averaged or coarse grained behavior of an assembly occurs with a simplified model for the individuals when they are coupled with a global set of inhibitory connections. At the same time we see an essential and important role for the details of spiking/burst-

ing neurons in achieving functional goals for these networks. The critical role of subthreshold oscillations in the inferior olive and in the olfactory processing as well as the importance of particular patterns of inhibitory coupling among these neurons join to produce accurate, robust, and reliable functional networks.

Acknowledgements

This paper is the result of numerous critical and often intense discussions with our colleagues at INLS (Al Selverston, Rob Elson, Attila Szucs, Joaquin Torres, Martin Falcke, Ramon Huerta, Nikolai Rulkov, Alexander Volkovskii, and Mike Maher), at Caltech (Gilles Laurent), and elsewhere (Rudolfo Llinas, John Rinzel, Gordon Shepherd, Terry Sejnowski, Maxim Bazhenov, and many others). Partial support for this work came from NSF grants NCR-9612250 and IBN-96334405. Mikhail Rabinovich acknowledges support from U.S. Department of Energy grant DE-FG03-96ER14592. Pablo Varona is supported by MEC. Henry Abarbanel is supported in part by U.S. Department of Energy grant DE-FG03-90ER14138 and in part by NSF grant NCR-9612250. Partial support for all authors also was received from the CIA/Office of Research and Development through project No. 98-F135000-000.

REFERENCES

Abarbanel, H.D.I., Huerta, R., Rabinovich, M.I., Rulkov, N.F., Rowat, P., and Selverston, A.I. (1996): 'Synchronized Action of Synaptically Coupled Chaotic Model Neurons', *Neural Computation, 8,* 1567.

Afraimovich, V.S., Verichev, N.N., and Rabinovich, M.I. (1986): *Inv. VUZ Radiofiz. RPQAEC, 29,* 795.

Bal, T., Nagy, F., and Moulins, M. (1998): 'The Pyloric Central Pattern Generator in Crustacea: a Set of Conditional Neuron Oscillators', *J. of Comparative Physiology A, 163,* 715.

Bazhenov, M., Huerta, R., Abarbanel, H.D.I., and Laurent, G. (1998): *Society for Neuroscience Abst., 24,* 359.6.

Berridge, M.J. (1998): 'Neuronal Calcium Signaling', *Neuron, 21,* 13.

Douglas, R.J., and Martin, K.C. (1991): 'Opening the Grey Box', *TINS, 14,* 7, 286.

Elson, R., Selverston, A.I., Huerta, R., Rabinovich, M.I., and Abarbanel, H.D.I. (1998): 'Synchronous Behavior of Two Coupled Biological Neurons', *Physical Review Letters, 81* (25), 5692.

Elson, R., Huerta, R., Abarbanel, H.D.I., Rabinovich, M.I., and Selverston, A.I. (1998): 'Dynamical Control of Irregular Bursting in an Identified Neuron of an Oscillatory Circuit', to appear.

Falcke, M., Huerta, R., Rabinovich, M.I., Abarbanel, H.D.I., Elson, R., and Selverston, A. (1998): 'Modeling Observed Chaotic Oscillations in Bursting Neurons: The Role of Calcium Dynamics and IP_3', submitted to *Biological Cybernetics.*

Getting, P.A. (1989): 'Emerging Principles Governing the Operation of Neural Networks', *Ann. Rev. Neurosci, 12,* 185.

Grillnet, S., Wallén, P., and Brodin, L. (1991): *Ann. Rev. Neurosci, 14,* 169.

Heagy, J.F., Pecora, L.M., and Carrol, T.L. (1994): *Physical Review, E 50,* 1874.

Hindmarsh, J.L., and Rose, R.M. (1984): *Proc. R. Soc. Lond., B 221,* 87.

Hodgkin, A.L., and Huxley, A.F. (1952): 'A Quantitative Description of Membrane Current and its Application to Conduction and Excitation in Nerve', *J. Physiol., 117,* 500.

Huerta, R., Rabinovich, M.I., Abarbanel, H.D.I., and Bazhenov, M. (1997): *Phys. Rev., E 55,* R2108.

Koch, C., and Segev, I. (1998): *Methods in Neuronal Modeling* (MIT Press).

Kristan, W.B. (1980): 'Generation of Rhythmic Motor Patterns', in *Information Processing in the Nervous System* (Raven Press, New York).

Laurent, G., Wehr, M., and Davidowitz, H. (1996): 'Temporal Representation of Odors in an Olfactory Network', *The Journal of Neuroscience, 16* (2), 3837.

Li, Y.-X., Keizer, J., Stojilkovic, S.S., and Rinzel, J. (1995): 'Ca^{2+} Excitability of the ER Membrane: an Explanation for IP3-induced Ca^{2+} Oscillations', *Am. J. Physiol., 269* (5 Pt 1), C1079.

Li, Y.-X., Rinzel, J., and Stojilkovic, S.S. (1995): 'Spontaneous Electrical and Calcium Oscillations in Unstimulated Pituitary Gonadotrophs', *Biophysical Journal, 69,* 785.

Li, Y.-X., Stojilkovic, S.S., Keizer, J., and Rinzel, J. (1997): 'Sensing and Refilling Calcium Stores in a Excitable Cell', *Biophysical Journal, 72,* 1080.

Llinás, R., Lang, E.J., and Welsh, J.P. (1997): 'The Cerebellum, LTD, and Memory: Alternative Views', *Learning & Memory, 3,* 445.

Llinás, R., and Welsh, J.P. (1993): 'On the Cerebellum and Motor Learning', *Curr. Opin. Neurobiol., 3*, 958.

Mori, K., and Shepherd, G.M. (1994): 'Emerging Principles of Molecular Signal Processing by Mitral/Tufted Cells in the Olfactory Bulb', *Seminars in Cell Biology, 5*, 1, 65.

Otsu, H., Yamamoto, A., Maeda, N., Mikoshiba, K., and Tashiro, Y. (1990): 'Immunogold Localization of Inositol 1, 4, 5-Trisphosphate ($InsP_3$) Receptor in Mouse Cerebellar Purkinje Cells Using Three Monoclonal Antibodie', *Cell Structure and Function, 15*, 163.

Parker, T.S., and Chua, L.O. (1989): *Practical Numerical Algorithms for Chaotic Systems* (Springer, New York).

Pearce, T.C. (1997): 'Computational Parallels between the Biological Olfactory Pathway and its Analogue the Electronic Nose', *Biosystems, 41* (1), 43.

Pecora, L.M., and Carroll, T.L. (1990): *Phys. Rev. Lett., 64*, 821.

Prechtl, J.C., Cohen, L.B., Pesaran, B., Mitra, P.P., and Kleinfeld, D. (1997): *Proc. Natl. Acad. Sci. USA, 94*, 7621.

Rabinovich, M.I., and Abarbanel, H.D.I. (1998): 'The Role of Chaos in Neural Systems', *Neuroscience, 87* (1), 5.

Rabinovich, M.I., Huerta, R., Bazhenov, M., Kozlov, A.K., and Abarbanel, H.D.I. (1998): *Physical Review, E 58*, 6418.

Rabinovich, M.I., Varona, P., Torres, J.J., Huerta, R., and Abarbanel, H.D.I. (1999): *Physica A*, in press.

Richardson, K.A., Imhoff, T.T., Grigg, P., and Collins, J.J. (1998): 'Encoding Chaos in Neural Spike Trains', *Physical Review Letters, 80* (11), 2485.

Rolls, E.T., and Treves, A. (1998): *Neural Networks and Brain Function* (Oxford University Press).

Rulkov, N.F., and Volkovskii, A.R. (1993): *Phys. Lett., A 179*, 332.

Satoh, T., Ross, Ch. A., Villa, A., Supattapone, S., Pozzan, T., Snyder, S.H., and Meldolesi, J. (1990): 'The Inositol 1, 4, 5-Trisphosphate Receptor in Cerebellar Purkinje Cells: Quantitative Immunogold Labeling Reveals Concentration in an ER Subcompartment', *The Journal of Cell Biology, 111*, 615.

Selverston, A.I., and Moulins, M. (1987): *The Crustacean Stomatogastric Systems* (Springer-Verlag, Berlin).

Selverston, A.I., Pachin, Y.V., Arshavsky, Y.I., and Orlovsky, G.N. (1997): 'Shared Features of Invertebrate Central Pattern Generators', in *Neurons, Networks and Motor Behavior* (MIT Press).

Shevelev, I.A., Tsicalov, E.N., Gorbach, A.M., Budko, K.P., and Sharaev, G.A. (1992): *J. Neurosci. Methods., 46*, 49.

Steriade, M., McCormick, D.A., and Sejnowski, T.J. (1993): *Science, 262*, 679.

Ullstrom, M., Kotaleski, J.H., Tegner, J., Aurell, E., Grillner, S., and Lansner, A. (1998): 'Activity-dependent Modulation of Adaptation Produces a Constant Burst Proportion in a Model of the Lamprey Spinal Locomotor Generator', *Biological Cybernetics, 79* (1), 1.

Varona, P., Torres, J.J., Huerta, R., Rabinovich, M.I., and Abarbanel, H.D.I. (1999): 'Regulation Mechanisms in the Bursting Behavior of Central Pattern Generator Neurons: A Modeling Study', to be submitted to *J. Neurophysiology*.

Walton, P.D., Airey, J.A., Sutko, J.L., Beck, C.F., Mignery, G.A., Südhof, T.C., Deerinck, T.J., and Ellisman, M.H. (1991): 'Ryanodine and Inositol Trisphosphate Receptors Coexist in Avian Cerebellar Purkinje Neurons', *The Journal of Cell Biology, 113*, 1145.

Wang, X.-J. (1993): *Physica, D 6*, 263.

AVOIDING NUCLEAR WAR

WOLFGANG K.H. PANOFSKY

This study-week addresses factors interfering with sustainable development, both gradually and through potential catastrophes that can endanger the survival of civilization. Military developments contribute to both of these threats. At the present time military activities in the name of national security consume on the average about one-third of the governmental expenditure of each country, with the less developed countries spending an even larger fraction of their fragile incomes on military matters than do the advanced industrial nations.

The economics underlying military spending are quite different from those underlying spending for civilian needs. Expenditure on civilian needs such as food, shelter, cultural activities and so forth are designed to address a need and, once these needs begin to be met, demand is decreased at least to some finite extent. Even saturation of demand is attainable so that an equilibrium level of sustainable development can be reached.

Military expenditure has precisely the inverse dynamic. As one nation increases its spending for military purposes, other nations perceive an increased threat and thus their military needs appear greater. As a result, military expenditure does not move in the direction of saturating a need but, in fact, augments rather than decreases demand. The result is an instability which can produce regional or global arms races and which can also lead to an apparent urgency and priority for military expenditure which takes priority over many other vital environmental and human considerations.

Whenever there have been arms races, economic and conservation considerations have been pushed into the background, as is witnessed by the awesome ecological heritage of the Cold War. Here the clean-up costs of environmental excesses, committed by the two super-powers in many of their military programs, have frequently become larger than the cost of the activities which left this destructive heritage. Examples abound: the U.S.

and Russia are now mounting multi-billion dollar programs to get rid of seventy thousand tons of chemical warfare agents produced during the Cold War. For technical and safety reasons, the rate of *dismantling* nuclear weapons has been unable to match the previous rate of *build-up* of nuclear weapons during the Cold War by a substantial factor.

This instability inherent in weapons competition among nations leads to runaway solutions which can either be terminated by catastrophe brought about by economic exhaustion or by outright war. But hopefully there can be mutually agreed limitations of military activity through the process of what is called arms control. The mission of arms control is to increase national and international security and reduce the costs and risks of armaments. While arms control has achieved many past important but sporadic successes, its overall record has not remotely eliminated the threat to sustainable development and survival which weapons acquisition and deployment imply.

We must recognise that the vast majority of deaths – some tens of millions – caused by warfare since the end of World War II have not been brought about by the feared weapons of mass destruction but by conventional weapons and largely small arms conventional weapons at that. None of these are significantly affected by arms control agreements and the existing limited export controls are bypassed to varying degrees by illegal activities. These problems are enormous and a separate study-week could easily be dedicated to the large scale impact of military activities on the human condition. I will talk only on a narrow but exceedingly important part of this impact and this has to do with the threat of nuclear weapons to human civilization. Here the threat is sudden disaster rather than a continuous toll of deaths and economic disruption.

The Pontifical Academy of Sciences has been deeply concerned about the threat posed to humanity by nuclear weapons. The Academy sponsored a number of studies on this subject leading to the "Declaration on the Consequences of the Use of Nuclear Weapons", released in October 1981, and the very important "Declaration on Prevention of Nuclear War", released on 24 September 1982. This Declaration was far-reaching, and it is a sad testimony to the current state of affairs with regard to nuclear weapons that the large majority of the admonitions in that Declaration are just as important and largely unfulfilled today as they were in 1982. Because of the importance of that Declaration, I am appending it to this paper. The Pontifical Academy also carried out a study on the weaponization of space in January 1985. That study attested to the impossibility of defenses against nuclear weapons and the importance of limiting such defenses in order to preventing a further escalation of nuclear weapons.

The first two nuclear weapons exploded in wartime over Hiroshima and Nagasaki; and killed about one-quarter of a million people. During the Cold War the "action-reaction" between the Soviet Union and the U.S. blessed the world with a deployment of over 60,000 nuclear warheads of an average explosive power more than ten times larger than those detonated over Japan. This build-up, in retrospect, was a clearly insane consequence of the Cold War and much has been written about the origins of that insanity. Dominant among the reasons was that the world permitted nuclear weapons to become seen as *symbols* of power and political influence rather than as physical objects of mass destruction. Moreover, while a norm appears to have been accepted internationally since Hiroshima and Nagasaki – that nuclear weapons shall not be actually used ever again in anger – the exercise of that norm has focused on deterrence, that is to say the threat of the use of nuclear weapons to deter what one possessor of nuclear weapons would consider to be unacceptable conduct on the part of others.

The concept of deterrence has been beset by a lack of clarity and understanding as to what conduct by a potential opponent would be deterred by deployment of nuclear weapons. First, we have what has recently been called the "core mission", that is, the function of nuclear weapons as a deterrent to the threat or actual use of *nuclear* weapons by others. It is generally agreed that for this purpose a relatively small number of *nuclear* weapons would be adequate. However, the concept of deterrence has been "extended" to other missions. The concept of "extended deterrence" implies that the party possessing nuclear weapons would be willing to threaten their use or actually employ them should a potential adversary engage in conduct contravening the interests of the possessor of nuclear weapons, and this where the interpretation as to what range of conduct is to be deterred can be quite broad.

The most important application of extended deterrence was the concept introduced by the United States and NATO during the Cold War that nuclear weapons were to compensate for the perceived inferiority of the NATO powers measured by strength in conventional weaponry. Therefore a potential invasion by the Warsaw Pact States of Europe was to be counteracted by the use of nuclear weapons by NATO. At the same time the Soviet Union declared a policy of no-first-use, stating that it would not use nuclear weapons first, that is only in response to the use of nuclear weapons by NATO.

Paradoxically this situation has now been reversed. While NATO has persisted in declaring that they are not bound by a no-first-use of nuclear weapons declaration, the Russians have officially abandoned their no-first-use stance, basing that reversal on the now acknowledged conventional

weapons superiority of the NATO powers in Western Europe. At the same time two NATO powers (Germany and Canada) are now advocating that NATO adopt a no-first-use policy. Thus we see the spectacle of a major power subscribing to a no-first-use declaration as long as that power believed itself to be superior in conventional weapons armaments, but revoking that declaration as soon as that conviction has eroded.

In addition to applying deterrence to non-nuclear threats, "extension" also implies that States possessing nuclear weapons would be willing to use them first in the case that there were major threats to the security of an allied nation, whether that threat were nuclear or non-nuclear. Thus, as the doctrine of deterrence has gone beyond its "core mission" through extension to other military objectives, the nations have denied themselves any meaningful criterion by which they can judge "when is enough, enough". Thus a meaningful concept of "sufficiency", to use the term coined by Henry Kissinger, has eluded the possessors of nuclear weapons. Thus the nuclear weapons arms race has led to the vastly excessive deployments which I have cited.

It has often been claimed that the arms race between the Soviet Union and the United States has been "won" by the U.S. Yet *both* sides have made enormous economic sacrifices on the altar of nuclear weapons. A recent analysis has shown that if the entire cost of nuclear weapons is taken into account, including their production, the cost of delivery vehicles, the cost of antinuclear defenses, the cost of infrastructure, and the cost of environmental clean up engendered by over-zealous production, then the total cost has been about five times 10^{12} dollars, or five trillion dollars expressed in American terms; about 29 percent of all U.S. military spending since World War II. Presumably, costs in real terms on the Soviet side were similar. Yet one must be cautious in criticizing nuclear weapons on economic grounds. Even the enormous figures cited are a small fraction of the total cost of the Cold War which has been equivalent to over a hundred trillion dollars. Paradoxically, nuclear weapons are the cheapest way to kill large masses of people for the simple reason that the explosive energy which can be carried by a vehicle of a given size is multiplied about a million fold. Thus the chief danger posed by nuclear weapons is their potential for a vast human catastrophe rather than their economic burden. Clearly the demise of the Soviet Union has many roots, including the fundamental frailty of the system. But at the time of sollapse of the Soviet Union the military-industrial complex constituted about three-fourths of the Soviet economy. This large fraction accounts in part for the difficulties in achieving economic recovery in Russia.

Now the Cold War is over, and the basic rationale for deploying nuclear weapons, whatever its past validity, has disappeared. Even while the Cold War was in full swing efforts commenced, based on the abvious threat

to the survival of civilization which the nuclear build-up implied, to halt and hopefully reverse the nuclear arms race. These efforts resulted in numerous arms control agreements in the nuclear field, including SALTI and the ABM treaty, the subsequent START I and the as yet unratified START II reduction treaties, the treaty on the reduction of intermediate nuclear forces in Europe which eliminated a whole class of nuclear weapons, limits on the testing of nuclear weapons and several other agreements, some formally negotiated and some effected by "reciprocal unilateral moves". An example of the latter are the initiatives taken by Presidents Bush and Gorbachev in drastically reducing the numbers of tactical nuclear weapons on land and sea.

As a result of all these measures the total number of nuclear weapons deployed today has receded to about one-half of its Cold War peak. But this decrease is unjustifiably small considering the drastic changes which have occurred in the world. The potential scale of nuclear destruction remains almost as large as it was at the peak of the Cold War and a large proportion of nuclear weapons are still on "hair-trigger", that is to say they can be launched at extremely short notice. This situation persists, notwithstanding the fact that today a deliberate nuclear attack by any of the five nuclear weapons States – the U.S., Russia, the UK, France and China – on one another is exceedingly unlikely. Rather, the nuclear dangers which are the most threatening stem from two sources: the unauthorized or accidental launch of nuclear weapons and the proliferation of nuclear weapons to States not now possessing them.

Stemming the proliferation of weapons of war has never succeeded in the history of mankind. Whenever a new technology has emerged which can be applied both to constructive civilian purposes and to forge the tools of war, it has never been possible to restrict the applications of that technology to peaceful purposes only. We must be mindful of this historical fact in addressing the threat of the proliferation of nuclear weapons. In that respect, notwithstanding the historical precedent, we *must succeed* or otherwise the future of human civilization is indeed in grave danger. A very good beginning towards controlling the proliferation of nuclear weapons was made with the nuclear Non Proliferation Treaty (NPT) which was signed in 1968 and which last year was made into a treaty of indefinite duration.

The NPT is an unprecedented achievement. The vast majority of the nations of this world have agreed that their national security is served better by forswearing the possession of nuclear weapons than by their acquisition. Only five States, which are the States which possessed nuclear weapons at the time of the signature of the NPT, are designated as Nuclear Weapons States while all the remaining signatories, which include almost all the

remaining States in the world have subscribed to the treaty as Non Nuclear Weapons States. However, in view of these facts, the treaty is fundamentally discriminatory; therefore an important part of the agreement which underlies the NPT are the provisions designed to erase over time this discriminatory nature of the NPT.

Under the NPT agreement, the Non Nuclear Weapons States agree not to build or receive nuclear weapons, and the Nuclear Weapons States agree not to supply them to others. At the same time, in order to diminish the discriminatory impact of the NPT, the Nuclear Weapons States agree to furnish assistance to Non Nuclear Weapons States to pursue the peaceful application of nuclear energy, in particular nuclear power. The relevant treaty provision requires that Non Nuclear Weapons States, in the pursuit of nuclear power, submit potential nuclear power plants to inspections by the International Atomic Energy Agency to prevent the diversion of nuclear fissionable materials to military purposes. But, above all, the NPT contains a provision which requires that Nuclear Weapons States reduce their dependence in their international relations on nuclear weapons and work in good faith towards their eventual prohibition.

It is in relation to this latter provision that the Nuclear Weapons States are rightfully subject to severe criticism by the Non Nuclear Weapons States. It is this criticism which is most likely to be heavily voiced during the next review conference for the NPT, which is scheduled for late this year. It is expected that these criticisms will focus on two major issues: (1) the lack of change in United States and NATO nuclear weapons policy, even given the end of the Cold War and the renunciation by Russia of the former no-first-use policy of the Soviet Union, and (2) the lack of arms control progress by the Nuclear Weapons States during the last few years. These criticisms are well founded.

The United States carried out a major review of its nuclear policy in 1994, called the Nuclear Posture Review, and last year President Clinton issued a new order covering nuclear weapons policy. Both revisions reflected disappointingly small changes with regard to past policy. The Nuclear Posture Review proclaimed a "reduce and hedge" policy, meaning that the past downward thrust in strategic arms limitation agreements should be continued but that the United States would hedge by retaining a sufficient inventory of complete nuclear weapons to make it possible to add additional warheads to long-range missiles capable of carrying them. This hedge was to provide protection should Russia revert to a more aggressive posture. The recent Presidential Order withdrew the requirement that the United States and Russia should be able to fight a protracted nuclear war, however unrealistic such a horrendous possibility might have been, and that

the United States should be prepared only for a single nuclear exchange. Under these policies, the United States will continue to retain an enduring stockpile of nuclear weapons, near to 10,000 units. As discussed above, Russia renounced its no-first-use policy and is also continuing the deployment of very large numbers of tactical nuclear weapons, presumably for protection against potential enemies near its borders. Thus both countries envisage total nuclear weapons stockpiles of comparable magnitude. There remains, however, the serious question of whether Russia can actually retain as many nuclear weapons as projected and keep them safe and reliable.

This leads me to the further major risk of future catastrophe, which is the possibility of the unauthorized launching of nuclear weapons or accidental nuclear explosions. The early warning system of Russia has deteriorated in quality and its coverage has shrunk because some of the early warning radars of the former Soviet Union are no longer located within the territory of Russia. This has already led to a case of a severe mis-identification: a Norwegian experimental rocket launch was erroneously identified as a possible attack against Russia from U.S. submarines. This event added to the many false alarms of the past, which happily did not lead to any actual disasters since the command and control systems of both countries contain numerous protective layers to prevent disaster. But a future risk of a disastrous nuclear weapons accident remains, given the many problems facing the military systems in the world, in particular the Russian command and control structure.

The second criticism, which the nuclear weapons nations will have to face up to during the NPT conference, is based on the lack of progress in actual arms control. This deficiency is rooted in the current weaknesses in leadership in Russia and the United States, and on the poor relations between the executive and legislative branches of government in the two countries. As a result, ratification of the very recent START treaty – the START II Treaty, which is expected to reduce the number of deployed strategic nuclear weapons on both sides to 3,500 – has been stalled in the Russian legislature. This lack of action is based more on political than military considerations – START II would constrain American deployments much more than Russian deployments since, for economic reasons, Russia could almost certainly not even build up to the START II limits.

Similarly, ratification of the recently signed Comprehensive Test Ban Treaty, to which all Nuclear Weapons States are signatories and which was identified during the NPT negotiations and in the wording of its recent extension as a condition for a robust nonproliferation regime, has not been approved by the U.S. Senate, and has not even been released by the relevant committee for discussion. Thus in both countries the legislatures have

failed to act in the interest of arms control progress, and in this have been driven by internal and external political motivations.

Possibly the dominant reason for the lack of progress in nuclear weapons arms control is public apathy. The world is in a peculiar situation: nuclear weapons have not been used in war since 1945 over Japan, and a norm of non-use of nuclear weapons appears to have been prevalent since that time. Testing in the atmosphere ceased in 1963 and all nuclear weapons tests have been now discontinued, something broken only by the Pakistan-Indian underground events last year. Only a very few old people have ever seen the reality of a nuclear explosion and therefore to most policy-makers nuclear weapons are, in essence, theoretical objects. Thus the world's leaders and populations are concerned across the world with much more imminent and visible problems than the possibility of nuclear catastrophe. Yet the risk of such a catastrophe is far from zero and we cannot postpone action until such an event accurs.

What can be done to reduce the risk of nuclear catastrophe? There is no single clear path. Nuclear weapons cannot be un-invented! It is most important to revitalize the arms control process both in its qualitative and quantitative aspects. Qualitatively, operational practices must be modified to sharply decrease the risk of unauthorized or accidental use of nuclear weapons. To further this goal, deployed nuclear weapons should be de-alerted, that is to say a mandatory time delay should be imposed, both through mechanical and doctrinal efforts, between a decision to launch and the actual release of nuclear weapons after the end of the Cold War. There simply is no need to be able to act in a matter of minutes, yet many systems are still configured for such a hair-trigger response. De-alerting has been studied extensively. Some technical measures toward this goal are straightforward, such as not routinely carrying nuclear weapons in aircraft, making silo covers difficult to open rapidly, removing warheads from deployed land-based missiles, and so forth. Other measures are more subtle and more difficult to verify, such as dealerting missiles on board submarines.

Nuclear weapons are still stored and deployed in large numbers but their materials are ageing and the older generation of expert designers of such weapons is phasing out. The United States, and to a lesser extent the other Nuclear Weapons States, have initiated a Science Based Stockpile Stewardship Program whose goal is to preserve a cadre of nuclear weapons experts and to engage in a systematic process of surveillance of the nuclear weapons stockpile and to maintain the capability to re-manufacture components should they prove unreliable or even dangerous. It is too early to judge the extent to which these and similar programs will be adequate to reduce the nuclear weapons safety risk. Clearly drastic reductions of

nuclear weapons are needed, both as an arms control measure in its own right and to reduce the number of weapons which must be controlled to prevent accidental or unauthorized release. The worldwide apathy regarding nuclear weapons must be reversed to exert pressure on the Nuclear Weapons States to subordinate their internal and external political interests to their overriding obligations to reduce the role of nuclear weapons in accordance with their obligations under the Nuclear Non Proliferation treaty, and in the simple interest of humanity.

The total inventories of the U.S. and Russia are clearly vastly excessive if the only remaining legitimate mission of nuclear weapons is to deter the use of nuclear weapons by others. While, indeed, circumstances can be envisaged in which the use of nuclear weapons might be to the military advantage of one party, it is the clear obligation of all Nuclear Weapons States to tailor their military policies so that the need for nuclear weapons will not be extended to any purpose other than this "core mission". Under this assumption, a regime of progressive reductions and restraints on nuclear weapons should be implemented by Russia and the U.S., and that regime should encompass the other Nuclear Weapons States as soon as reductions in the numbers of nuclear weapons of these two States have made their inventories comparable to those of the other Nuclear Weapons States.

The world should now face the question of the total prohibition of nuclear weapons. I deliberately use the word "prohibition" rather than "elimination". The decision to prohibit nuclear weapons must be based on a balance of risks: the current dangers of nuclear weapon proliferation and the accidental or unauthorized explosion of nuclear weapons must be balanced against the risk that under an international norm prohibiting nuclear weapons there will still be small numbers of nuclear weapons retained by Nuclear Weapons States or clandestinely manufactured by national or even subnational groups. The time has come to place an evaluation of this balance of risk firmly on the international agenda and to examine the feasible paths which can lead to the prohibition of weapons.

These conclusions are consonant with the recommendations contained in the "Declaration on Prevention of Nuclear War" which the Pontifical Academy released on 24 September 1982. That Declaration contained an admonition that all nuclear nations should adopt the policy of no-first-use of nuclear weapons and intensify their efforts to reach verifiable agreements curbing the arms race. The Declaration also contained an admonition to take all practical measures to reduce the possibility of nuclear war by accident, miscalculation or irrational action, and to intensify efforts to prevent the further proliferation on nuclear weapons. The Declaration was addressed to national and religious leaders, to scientists, and to people everywhere. I

hope that this paper will remind those assembled here that, while some progress along the lines indicated has been made, both the rate and the actual achievements are woefully insufficient in reducing nuclear weapons as a major, if not the largest, threat to sustainable development and survival.

Let me reiterate again that the grave dangers posed by nuclear weapons are only a small part of the larger problem posed to sustainable development by the unacceptably large diversion worldwide of economic and human resources from productive civilian activities to military pursuits, which each country undertakes in the name of its national security but which in reality contribute to worldwide insecurity.

APPENDIX

DECLARATION ON PREVENTION OF NUCLEAR WAR*

by an assembly of Presidents of Scientific Academies
and other scientists from all over the world
convened by the Pontifical Academy of Sciences

September 23-24, 1982

I. Throughout its history, humankind has been confronted with war, but since 1945 the nature of warfare has changed so profoundly that the future of the human race, of generations yet unborn, is imperiled. At the same time, mutual contacts and means of understanding between peoples of the world have been increasing. This is why the yearning for peace is now stronger than ever. Mankind is confronted today with a threat unprecedented in history, arising from the massive and competitive accumulation of nuclear weapons. The existing arsenals, if employed in a major war, could result in the immediate deaths of many hundreds of millions of people, and of untold millions more later through a variety of after-effects. For the first time, it is possible to cause damage on such a catastrophic scale as to wipe out a large part of civilisation and to endanger its very survival. The large-scale use of such weapons could trigger major and irreversible ecological and genetic changes, whose limits cannot be predicted.

Science can offer the world no real defense against the consequences of nuclear war. There is no prospect of making defenses sufficiently effective to protect cities since even a single penetrating nuclear weapon can cause massive destruction. There is no prospect that the mass of the population could be protected against a major nuclear attack or that devastation of the cultural, economic and industrial base of society could be prevented. The breakdown of social organization, and the magnitude of casualties, will be so large that no medical system can be expected to cope with more than a minute fraction of the victims.

There are now some 50,000 nuclear weapons, some of which have yields a thousand times greater than the bomb that destroyed Hiroshima. The total explosive content of these weapons is equivalent to a million Hiroshima bombs, which corresponds to a yield of some three tons of TNT for every person on earth. Yet these stockpiles continue to grow. Moreover, we face the increasing danger that many additional countries will acquire nuclear weapons or develop the capability of producing them.

There is today an almost continuous range of explosive power from the smallest battlefield nuclear weapons to the most destructive megaton warhead. Nuclear weapons are regarded not only as a deterrent, but there are plans for their tactical use and use in a general war under so-called controlled conditions. The immense

* Presented to His Holiness the Pope by an assembly of Presidents of scientific Academies and other scientists from all over the world convened by the Pontifical Academy of Sciences.

and increasing stockpiles of nuclear weapons, and their broad dispersal in the armed forces, increase the probability of their being used through accident or miscalculation in times of heightened political or military tension. The risk is very great that any utilization of nuclear weapons, however limited, would escalate to general nuclear war.

The world situation has deteriorated. Mistrust and suspicion between nations have grown. There is a breakdown of serious dialogue between the East and West and between North and South. Serious inequities among nations and within nations, shortsighted national or partisan ambitions, and lust for power are the seeds of conflict which may lead to general and nuclear warfare. The scandal of poverty, hunger, and degradation is in itself becoming an increasing threat to peace. There appears to be a growing fatalistic acceptance that war is inevitable and that wars will be fought with nuclear weapons. In any such war there will be no winners.

Not only the potentialities of nuclear weapons, but also those of chemical, biological and even conventional weapons are increasing by the steady accumulation of new knowledge. It is therefore to be expected that also the means of non-nuclear war, as horrible as they already are, will become more destructive if nothing is done to prevent it. Human wisdom, however, remains comparatively limited, in dramatic contrast with the apparently inexorable growth of the power of destruction. It is the duty of scientists to help prevent the perversion of their achievements and to stress that the future of mankind depends upon the acceptance by all nations of moral principles transcending all other considerations. Recognizing the natural rights of humans to survive and to live in dignity, science must be used to assist humankind towards a life of fulfillment and peace.

Considering these overwhelming dangers that confront all of us, it is the duty of every person of good will to face this threat. All disputes that we are concerned with today, including political, economic, ideological and religious ones, which are not be undervalued, seem to lose their urgency compared to the hazards of nuclear war. It is imperative to reduce distrust and to increase hope and confidence through a succession of steps to curb the development, production, testing and deployment of nuclear weapons systems, and to reduce them to substantially lower levels with the ultimate hope of their complete elimination.

To avoid wars and achieve a meaningful peace, not only the power of intelligence are needed, but also the powers of ethics, morality and conviction.

The catastrophe of nuclear war can and must be prevented. Leaders and governments have a grave responsibility to fulfill in this regard. But it is human kind as a whole which must act for its survival. This is the greatest moral issue that humanity has ever faced, and there is no time to be lost.

II. In view of these threats of global nuclear catastrophe, we declare:

— Nuclear weapons are fundamental different conventional weapons. They must not be regarded as acceptable instruments of warfare. Nuclear warfare would be a crime against humanity.

— It is of utmost importance that there be no armed conflict between nuclear powers because of the danger that nuclear weapons would be used.

— The use of force anywhere as a method of settling international conflicts entails the risk of military confrontation of nuclear powers.

— The proliferation of nuclear weapons to additional countries seriously increases the risk of nuclear war and could lead to nuclear terrorism.

— The current arms race increases the risk of nuclear war. The race must be stopped, the development of new more destructive weapons must be curbed, and nuclear forces must be reduced, with the ultimate goal of complete nuclear disarmament. The sole purpose of nuclear weapons, as long as they exist, must be to deter nuclear war.

III. Recognizing that excessive conventional forces increase mistrust and could lead to confrontation with the risk of nuclear war, and that all differences and territorial disputes should be resolved by negotiation, arbitration or other peaceful means, we call upon all nations:

— Never to be the first to use nuclear weapons.

— To seek termination of hostilities immediately in the appalling event that nuclear weapons are ever used.

— To renew and increase efforts to reach verifiable agreements curbing the arms race and reducing the numbers of nuclear weapons and delivery systems. These agreements should be monitored by the most effective technical means. Political differences or territorial disputes must not be allowed to interfere with this objective.

— To find more effective ways and means to prevent the further proliferation of nuclear weapons. The nuclear powers, and in particular the superpowers, have a special obligation to set an example in reducing armaments and to create a climate conducive to non-proliferation. Moreover, all nations have the duty to prevent the diversion of peaceful uses of nuclear energy to the proliferation of nuclear weapons.

— To take all practical measures that reduce the possibility of nuclear war by accident, miscalculation or irrational action.

— To continue to observe existing arms limitation agreements while seeking to negotiate broader and more effective agreements.

IV. Finally we appeal:

1) To national leaders, to take the initiative in seeking steps to reduce the risk of nuclear war, looking beyond narrow concerns for national advantage; and to eschew military conflict as a means of resolving disputes.

2) To scientists, to use their creativity for the betterment of human life, and to apply their ingenuity in exploring means of avoiding nuclear war and developing practical methods of arms control.

3) To religious leaders and other custodians of moral principles, to proclaim forcefully and persistently the grave human issues at stake so that these are fully understood and appreciated by society.

4) To people everywhere, to reaffirm their faith in the destiny of humankind, to insist that the avoidance of war is a common responsibility, to combat the belief that nuclear conflict is unavoidable, and to labor unceasingly towards ensuring the future of generations to come.

E. Amaldi (Italy)
I. Badran (Egypt)
A. Balevski (Bulgaria)
D. Baltimore (USA)
A. Bekoe (ICSU)
F. Benvenuti (Italy)
C. Bernhard (Sweden)
O. Bikov (USSR)
B. Bilinski (Poland)
C. Chagas (Brazil)
E. De Giorgi (Italy)
B. Dinkov (Bulgaria)
G. Hambraeus (Sweden)
T. Hesburgh (USA)
H. Hiatt (USA)
D. Hodgkin (PUGWASH)
S. Hsieh (Taipei)
A. Huxley (UK)
S. Iijima (Japan)
S. Isaev (USSR)
P. Jacquinot (France)
W. Kalweit (GDR)
M. Kazi (Pakistan)
S. Keeny (USA)
K. Komarek (Austria)
F. Konig (Austria)
J. Labarbe (Belgium)
J. Lejeune (France)
L. Laprince-Ringuet (France)
R. Levi Montalcini (Italy)
M. Lora-Tamayo (Spain)
T. Malone (USA)
G. Marini-Bettolo (Italy)
S. Mascarenhas (Brazil)
M. Menon (India)
R. Peierls (UK)
M. Peixoto (Brazil)
J. Peters (Belgium)
G. Porter (UK)
F. Press (USA)
G. Puppi (Italy)
B. Rifai (Indonesia)
W. Rosenblith (USA)
P. Rossano (Italy)
P. Rudomin (Mexico)
B. Rysavy (Czechoslovakia)
I. Saavedra (Chile)
V. Sardi (Venezuela)
T. Shin (Korea)
E. Simpson (South Africa)
J. Sirotkovic (Yugoslavia)
L. Sosnovski (Poland)
A. Stoppani (Argentina)
J. Szentagothai (Hungary)
S. Tanneberger (GDR)
C. Townes (USA)
E. Velikhov (USSR)
W. Watts (Ireland)
V. Weisskopf (USA)
K. Weizacker (FRG)

GREED AND IGNORANCE: MOTIVATIONS AND ILLUSTRATIONS OF THE QUANTIFICATION OF MAJOR RISKS

ELISABETH PATÉ-CORNELL

1. Risks and Bottlenecks

Major risks and scientific limitations

The great risks that weigh over mankind include epidemics, wars and weapons of mass destruction, demographic and social explosions, natural catastrophes, and massive climatic upsets. Science and engineering can address these issues, first by doing no harm, then by providing a better understanding of hazardous phenomena, and finally by designing remedial policies based on engineering or medical advances. Obstacles to the implementation of risk management solutions are thus of three types: scientific, technical and human.

Critical *scientific* bottlenecks include limited knowledge about cancer and viral mechanisms, the climate – in particular, the role of the clouds and of the oceans –, the earth's tectonics, and also social phenomena such as the rise and resolution of open or latent conflicts. Critical *technical* bottlenecks as defined in engineering, are the weak points of a system that need to be addressed in priority to prevent a disaster. The search for these technical bottlenecks involves an analysis of interactions among the different technical elements of a system, and of the probability of system failure as a function of the chances of component failures. In recent years, these technical risk analyses have been extended to include system weaknesses caused by management failures and human errors. (e.g., Paté-Cornell and Murphy, 1996) This extension of the method which allowed us to identify a wide spectrum of correctable weaknesses, is described in the examples of section two of this paper. For example, in a study of the tiles of the space shuttle and of the risk of an accident caused by a failure of the orbiter heat shield,

we found that the main problem was the strength of the tiles' bond; but the real system weakness at that time was a management failure to reduce unnecessary time pressures.

Therefore, the search for effective risk mitigation solutions requires not only a serious investment in fundamental research (Alfimov *et al.*, 1997), but also the development of sophisticated methods of analysis involving technical, human and organizational factors to identify system weak points and permit the crafting of effective risk management policies. Implementing these policies then demands a commitment of limited resources. It also often requires making the best use of incomplete or imperfect knowledge when risk mitigation measures must be taken before full information is available.

Human limitations: greed and ignorance

This commitment of resources and the need to act based on incomplete information require overcoming some basic human propensities. One is to consume now rather than later, often ignoring the limitations of the earth's resources and the effect of national policies on other people, now and in the future. Another is the human tendency, unwillingness or inability to recognize the limits of knowledge at a given time. When forced to depart from the comfort of scientific certitudes, summarizing incomplete information and processing uncertainties in a logical manner is a difficult exercise that often exceeds the capacity of the human mind. As it is described further, this synthesis can be performed systematically through probabilistic reasoning. Some of the bottlenecks in the design and implementation of sound policies to address problems of survivability and sustainability may thus be scientific and economic. But the most fundamental ones are human in nature. They are greed and ignorance.

Searching for system weaknesses

In addition to funding basic scientific, medical and engineering research, we thus need to invest in the development of analytical tools that allow addressing these shortcomings, whether scientific, technical or human. These tools include methods of risk analysis, based on logic and probability to characterize uncertainties, and economic analysis to quantify the consequences of different hazard scenarios. Probabilistic methods provide decision support for the design of risk mitigation policies accounting for limitations of both resources and knowledge. In particular, they allow identification of system weaknesses both technical and human, and evalua-

tion of the benefits of a spectrum of remedial options. Social science can play a major part in understanding and addressing human problems at the scale of cities as well as nations. I do not attempt here to address explicitly its role in the mitigation of major social, economic and political threats. Yet, population explosion, for instance, is one of the fundamental hazards in the world, and population control can be effective only if it is harmoniously integrated into the life of the different societies. Instead, I focus here on the physical and natural sciences, although I use some elements of social sciences to assess the contribution of human and organizational problems to the risk of physical system failures.

Policy decisions under uncertainty

Global climate change is a good example of a potential threat that requires serious consideration because it has the potential to cause massive shifts from the current state of the world. The scientific experts, however, are divided. Large uncertainties remain about the fundamental mechanisms involved, the role of human actions (as opposed to natural trends), the results that can be expected from different public policies, and the effects on the world economy of both the phenomenon and the costs of proposed solutions (Paté-Cornell, 1996a). Failure to address the problem in time may precipitate what can turn out to be an irreversible catastrophe. Yet, investing massive amounts in corrective actions without considering uncertainties may result in great expenses and few benefits. Before making costly policy decisions, a careful analysis of the uncertainties may be helpful, first to decide where to invest in scientific research and how to characterize the state of knowledge at a given time. That is not to say that quantifying the risks of global climate change is easy: it is extremely complex because of the great number of factors, parameters and feedback mechanisms involved. But ignoring the possibility and the probabilities of alternative hypotheses can perpetuate errors and lead to dangerous policy choices.

The next question is how to use that imperfect information for public policy decisions, an issue that involves individual or collective preferences and decision criteria (Arrow, 1963). Risk analysis can only provide the probabilities and the magnitudes of the consequences of different courses of action. Society must then make decisions based on that information, but these decisions will not be handed out to nations by science or engineering. Policy choices are the products of a social process and are only as good as this process (Paté-Cornell, 1983a). This paper is dedicated to the question of bringing relevant information into this policy decision process. The *use* of that information and the quality of the decision depend on the organiza-

tion and political system of the concerned group, and on its ability to involve, whenever possible, the people whose fate is at stake. Therefore, computations of existing risks and of the consequences of risk mitigation policies do not guarantee that greed will not prevail and that ignorance will be dispelled. Even if they are correct, the results only provide information that *can* be useful for a sensible allocation of resources, namely, what is known at decision time and what are the anticipated effects of different courses of action on us and on others. The ultimate effect will also depend on the values, the possibilities and the wisdom of those who will design and implement the corresponding policies in the future.

In the remainder of this paper, I present first several examples (technical and medical) of risk analysis studies that have allowed in practice, identification and resolution of technical and human problems in the search for risk mitigation solutions. I then address the realities of greed and ignorance, and their natural influence on policies and decisions that can be critical to mankind. The focus is on the acquisition and the gathering of information that is necessary (but not sufficient) to address these shortcomings. Next, I examine the role of science and engineering – in particular the use of engineering tools and probabilistic methods – in the management of major risks. I describe the nature and characterization of risk uncertainties (including the probabilistic treatment of expert disagreements) and the value of this information in the context of decision making. Finally, and without claim of exhaustivity, I discuss some of the limitations of probabilistic methods and of the moral and ethical problems that they may raise.

2. Examples of Risk Analyses Based on Probability

Extension of the risk analysis model to include human and organizational factors

Most probabilistic risk analysis (PRA) models focus on engineering and scientific features of systems and technical component failures; but they do not capture the effects of organizational and human factors that are often the root causes of these failures. Yet, addressing these management problems may be the simplest and the most cost-effective way to reduce the risk. To assess the risk contribution of these factors, we needed first to extend the essentially technical PRA model to account for their effects.

In many PRAs, the first step is to identify by Boolean (logical) analysis of a system's functions, the set of failure modes and failure scenarios, i.e., the conjunctions of events, phenomena and parameter values that can lead to a failure. The second step is to compute the probability of system failure,

given the performance of its components and the probabilities of external events that can affect several components simultaneously.

To integrate human and organizational factors into the assessment of the risks, we construct first, a technical PRA model. For each of the events or state variables of that model, we then identify the human decisions and actions (including errors) that affect the technical variables, and for each decision and action, the management factors that cause or influence them (Paté-Cornell and Murphy, 1996). This SAM model (System-Action-Management) involves marginal and conditional probabilities to capture the dependencies among the different. It permits identifying not only technical weaknesses, but also the human errors that contribute most to the risk and can be avoided, sometimes at a much lower cost than the reinforcement of physical components that would yield the same risk reduction benefits. This extension therefore, enlarges the scope of options and permits a better allocation of risk mitigation resources.

Detection of technical and human weaknesses: the case of the tiles of the space shuttle

Following the Challenger accident in 1986, NASA's managers and astronauts were concerned, among other problems, at the potential weaknesses of the space shuttle heat shield. One potential weakness was that of the bond between the black tiles that protect the underside of the orbiter and its aluminum surface. Also, a misalignment of the tiles can cause an early shift from laminar to turbulent flow at reentry in the atmosphere, thus increasing the overall heat load on the orbiter. But we found that the root causes of these technical weaknesses were essentially human and organizational. For instance, under time pressures, NASA did not consider any priorities among the tiles that had to be fixed or replaced. Also, in order to meet deadlines, the technicians were taking shortcuts that could be damaging to the tile system. The question was thus to identify and correct these management and labor problems, which as we show below, allowed NASA to reduce by 70% this component of the shuttle risk at very little cost.

In that study (Paté-Cornell and Fischbeck, 1994), we quantified the risks of losing a US space shuttle due to different failure modes of the black tiles. We estimated the contribution of human and management errors to the failure risk, and we made a number of recommendations to NASA to improve the maintenance and therefore, the safety, of the thermal protection system. First, we created a risk analysis model. We identified and quantified the risk factors (heat loads, aerodynamic forces, density of debris hits, and criticality of the subsystems under the orbiter's skin) for 33 zones of the orbiter's surface that we had defined according to these factors. The result

was a risk-based map of the orbiter showing in which zones, each tile contributed most to the probability of a shuttle accident due to tile failure.

We then investigated on the site (Kennedy Space Center) the types of errors that the tile technicians made, the shortcuts that they took and the reasons for these shortcuts in order to assess their effects on the failure risk and to make management recommendations. We found first, that the tiles contributed about 10% of the risk of a shuttle accident, and second, that all tiles are not equally critical: 15% of them account for 85% of the risk. Third, we discovered that under tight schedules, people were creative: one technician was found to spit in the glue to make it cure faster. The problem is that the catalytic chemical reaction that takes place at that time can revert itself earlier if water is added to the bond. This happened in part because of questionable time pressures. For example, the work schedule was sometimes defined in terms of number of tiles per day instead of per week, which would have left more flexibility in the organization of the work. We also found that the turnover among tile technicians was high and probably affected the quality of their work. This was apparently due to a minor discrepancy in salary – inherited from the U.S. Department of Defense – between the tile technicians and the machinists or electricians. Finally, we pointed out that small pieces of insulation material that could de-bond at takeoff from the surface of the external tank were a substantial element of the tile failure risk.

Once we had identified these system weaknesses, we made a number of recommendations that were generally implemented by NASA. Some of them were technical (e.g., reinforce the bonding of the insulation of the external tank). Most of them, however, were managerial in essence. They included, for example, focusing the test of 10% of the tiles before launch on the most risk-critical zones, increasing the tile technician salaries to retain them after training, and decreasing the time pressures on the tiles maintenance crews by giving them more flexibility. We computed (coarsely) the benefits of some of these risk mitigation measures, and showed that at very low cost, one could immediately reduce by 70% the contribution of the tiles to the risk of shuttle loss (about 10%). The key was to choose risk mitigation measures that corresponded to the weakest physical links (the bonds), the most frequent and serious human errors (short cuts and misalignment of the tiles), and the most effective management changes (keep trained people and recognize priorities in the work).

Detection of personnel and management weaknesses: the case of anesthesia patient risks

The risks of death of serious brain damage to patients subjected to anesthesia in modern western hospitals is in the order of one in ten thou-

sand. Isolating statistically the contribution of accident root causes, both technical and human, is impossible at this time because the data are not available. To identify the weak links of the anesthesia system requires a systematic analysis of all classes of accident scenarios, including human errors in the detection and interpretation of signals of problems. Therefore, the same method of probabilistic risk analysis extended to include human management factors was useful in this case. The challenge was to perform a dynamic analysis of accident scenarios based on all available evidence, including frequencies of minor incidents and expert opinions. Some of the most helpful among these experts were experienced operating room nurses. Surprisingly, the problems that had motivated our work (e.g., substance abuse among the practitioners) turned out to be minor risk contributors. More mundane issues of periodic recertification, resident supervision and practitioners' fatigue appeared to be the more salient weaknesses of the system and should be addressed in priority by hospital administrations.

In that 1996 study, our objective was thus to find the major root causes of anesthesia accidents and the most effective means of improving patient safety (Paté-Cornell *et al.*, 1996). We developed a patient risk model, we estimated the probabilities and the effects on this risk of personnel problems among anesthesiologists, and the potential risk reduction benefits of a number of management options such as the periodic retraining of practitioners on anesthesia simulators. The model that we used was a variation of the SAM framework described above. Following different "initiating events", the critical variable is the time needed to detect, diagnose and address different situations, for example, a patient allergic reaction to an anesthetic drug. We had statistical data regarding occurrences of initiating events and fatal accidents. We used expert opinions to model full accident sequences (intermediate events), and because we had these statistics at both ends of the model, we had the possibility of a "reality check". Second, with the help of experts, we divided the population of anesthesiologists according to ten types of personal problems that can affect their performance, for instance, severe distraction or extreme fatigue. We then evaluated the probability that they experience a specific problem during any given operation. We also assessed the effects of each of these personnel problems on the time that it takes for the anesthetist to properly detect and react to an incident given his or her state of competence and alertness. We could then estimate the effects of each type of anesthetist problem on a patient's risk of death or severe brain damage. Finally, we identified a number of management options (e.g., re-certify all anesthetists every five years), and we assessed their effects on the probabilities of personnel problems, and therefore, on the patient risk.

This analysis allowed us to rank the measures that we had identified according to their anticipated benefits, and to make a number of recommendations, which may or may not be implemented by hospital administrations depending on the costs and popularity of these policies. We were surprised to find that the dominant problems were much more common and "closer to home" than the more visible substance abuse issues that had motivated the study. We found that the most effective measures probably included a periodic retraining of the practitioners to increase the safety of those who do not operate frequently and may thus have forgotten how to respond to rare incidents. They also included closer supervision of residents and trainees than what the practice (if not the rule) has sometimes been. One of the most critical links of the safe resolution of an incident is the quick detection of a problem. In that respect, it became clear that the operating room nurses were playing a major role because they have the experience that allows them to observe and detect problems, sometimes faster than some attending physicians. They were also a critical source of information in our study.

Design and optimization of warning systems

Well-conceived warning systems are one of the most effective ways to mitigate large-scale risks, and probabilistic methods can be used to maximize their efficiency. These systems however are seldom perfect because they can both issue false alarms and miss event signals. One weakness may therefore be the "cry-wolf" effect in which people cease to respond after too many false alerts. On the other hand, either because a system is not sensitive enough or because it involves a chain of components in which transmission failures may occur, it can fail to issue a timely warning. Again, extending the risk analysis method was necessary to blend technical and human aspects of the problem in order to optimize the design and operations of such a system. What we found is that the optimal sensitivity depends on many factors and cannot be guessed *a priori*. The first question is to identify what to monitor and what signals to look for. The second is what to do during the lead time. The third is often to decide what is an appropriate warning threshold, i.e., one that does not make the system too sensitive but provides enough time for action.

In that study (Paté-Cornell, 1986a), we designed a general model that yields an optimal threshold of alert and is applicable to warnings of droughts as well as fire alarms. This threshold is based on the characteristics of the phenomenon monitored (frequencies and severity levels), the human memory of past system performance (and willingness to respond to

a new signal), and the effectiveness of risk mitigation actions that can be taken following a warning and given the lead time.

Consider, for example, a potential hazard that can be represented by a continuous stochastic process such as the level of water of a river, or the density of smoke (or other suspended particles) in the air. Its characteristics can be used to design a warning system that triggers an alarm when it exceeds a warning threshold, assuming that substantial losses and damage would occur above a higher "critical" threshold (see figure 1).

Most such systems are not perfect and can generate errors. They can fail to alert if the system does not function, or if in practice, the lead time is insufficient to respond effectively (Type I errors). They can cause false alarms when the warning threshold is exceeded but the critical threshold is not reached, or simply if the system malfunctions independently from the underlying process (Type II errors). Therefore, an optimal alert threshold on the one hand, does not generate so many false alerts that it discourages response (the "cry wolf" effect), and on the other hand, provides enough lead time for people to react and to protect themselves. The optimization model that maximizes the benefits (avoided losses) of measures taken during the lead time thus involves three different sub-models (Paté-Cornell, 1996a).

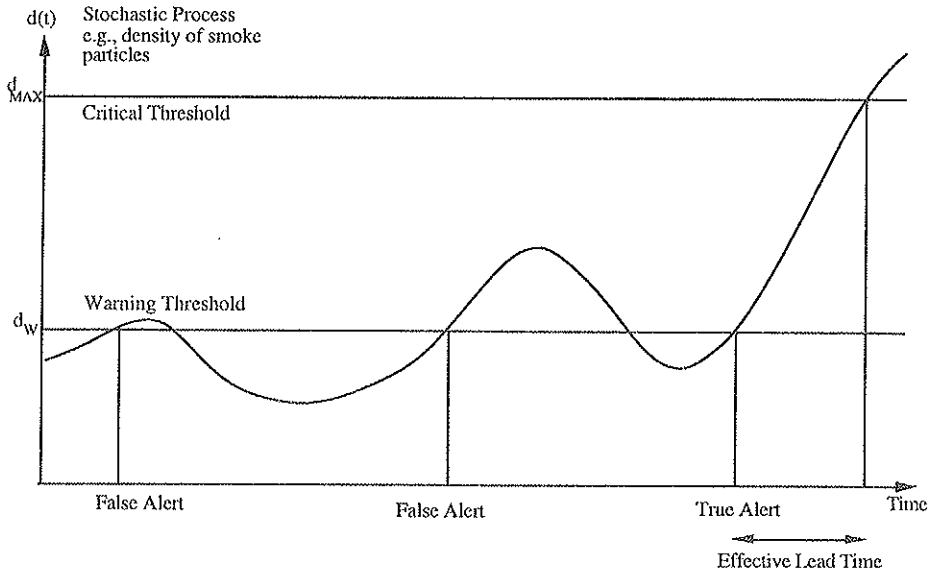


Fig. 1. Mathematical model of a warning system for a continuous stochastic process (Paté-Cornell, 1996a).

The first one is the description of the underlying physical phenomenon as a stochastic process. It includes the upcrossing rates of different alert thresholds (how often each of them is exceeded), and the rate of deterioration of the phenomenon following a true alert. This rate of deterioration, which can be represented by the mean slope of the curve between the warning and the critical thresholds, provides an estimate of the lead time available for actions such as evacuation.

The second model represents the effect of human memory of the system's past performance on people's willingness to respond to a new warning. Systems that emit too many false alerts are often ignored or turned off unless they monitor something that can turn into a truly catastrophic situation. This has been true of fire alarms, red lights indicative of engine fires in military airplane cockpits, and patient monitors in operating rooms. The result of this behavioral model is a rate of response to a new alert given a previous pattern of true and false alerts.

Finally, the benefits of a warning system are determined by the actions that can be taken during the lead time that the system provides, i.e., the time elapsed between upcrossings of the warning and the critical thresholds. The third model thus represents the options available following an alert and during the lead time, and their effectiveness in reducing the potential losses.

This probabilistic model – and several of its variations – have been used to address the tradeoff between the rates of Type I and Type II errors for several kinds of warning systems and response policies. It can also be used to identify critical weaknesses of a warning system and to reduce the rate of errors that need to be addressed in priority. We found that the maximum benefits from such systems seldom correspond to automatic response to the slightest indication of possible problems. Indeed, there is generally an optimal system sensitivity that depends on the memory effect of past alerts and on the effectiveness of actions taken during the available lead time. Applications have included fire alarms and monitoring of dams (Paté-Cornell, 1986b). Similar models can be used to identify, detect and respond to signals of famines or impending financial crises.

Estimating and mitigating risks when several hypotheses exist: use of the Bayesian method

Scientists can disagree for a long time before reaching a conclusion about a phenomenon that constitutes a threat (e.g., food contamination). Yet, risk mitigation policy decisions sometimes have to be made quickly, before complete information is available. At that stage, there are often a

number of competing hypotheses that provide different risk results, for example, different probabilities of occurrence of a catastrophic event. Furthermore, each new piece of evidence, observation and data is generally ambiguous: it does not automatically imply that a particular hypothesis is unquestionably correct, or whether the event of interest will definitely occur or not. The challenge in such a case is first, for the proponents of the different theories to accept the possibility of other hypotheses. The difficulty, for the decision maker, is to blend contradictory assumptions in a sound policy decision. Mathematical models that consider and update the probabilities of the different hypotheses and of the event of interest given each hypothesis, allow overcoming the obstacle of expert dissent in critical situations.

A Bayesian reasoning, which is discussed in details further in this paper, allows computation of the overall risk. It is based first, on the probability of each hypothesis given all available evidence (including the new observation, statistics or signals) and second, on the probability of the event given each hypothesis. We developed a model that allows for the computation and display not only of the risk itself, but also of the uncertainties about it that result from the existence and the probability of the different hypotheses (Paté-Cornell and Fischbeck, 1995).

Consider, for example, a situation where two fundamental hypotheses have been identified and constitute competing bases for the occurrence of a catastrophic event or phenomenon. For each of them, there is an associated parameter value, for example, the probability of this event given the hypothesis. As new pieces of evidence and data are integrated in the process, the probabilities of the hypotheses and of the event given each hypothesis can be updated so that a decision can be supported by the best representation of the uncertainties at that time. The model provides the mean probability of the event (comparable, for example, to a mean future frequency) but also a display of the uncertainties about this event probability, given the existence and the probabilities of the two underlying hypotheses. This simplified formulation is similar to cases such as global warming where several hypotheses are competing in the scientific world, but may have to be reconciled sooner rather than later in the world of policy decisions.

The illustration that follows shows how, in principle, one can provide a decision maker with a full description of these uncertainties, which may be critical, for instance, when the decision of interest is a matter of survival. The question was to assess the probability of a catastrophic event based on observed signals (or data) from a monitoring or observation system subjected to both Type I errors (false negatives) and Type II errors (false positives). Assume that the decision problem is whether and how to respond to

new data, given the evidence and before the event occurs, accounting for the possibility of observation errors and multiple hypotheses. The objective of the risk analyst is thus to derive, from this partial information, the probability of the event at the time when a decision needs to be made and based on available information. Figure 2 shows the process of updating of the probability of the event of interest for two hypotheses H_1 and H_2, both with 0.5 prior probabilities before new data are acquired. These assumptions could represent, for instance competing theories from two experts whom the decision maker (who assesses the probabilities) equally trusts.

Assume for the sake of this illustration, that the prior probabilities of the event, before new evidence is available and given each hypothesis, are 0.2 given H_1 and 0.3 given H_2, thus yielding a 0.25 mean prior probability of the event. After updating based on the new data, the probabilities of the hypotheses are 0.42 and 0.58 respectively. These figures are computed based on some assumed probabilities of observing the data given that the event will and will not occur respectively. The conditional posterior probabilities of the event are then 0.83 given H_1 and 0.89 given H_2 yielding a pos-

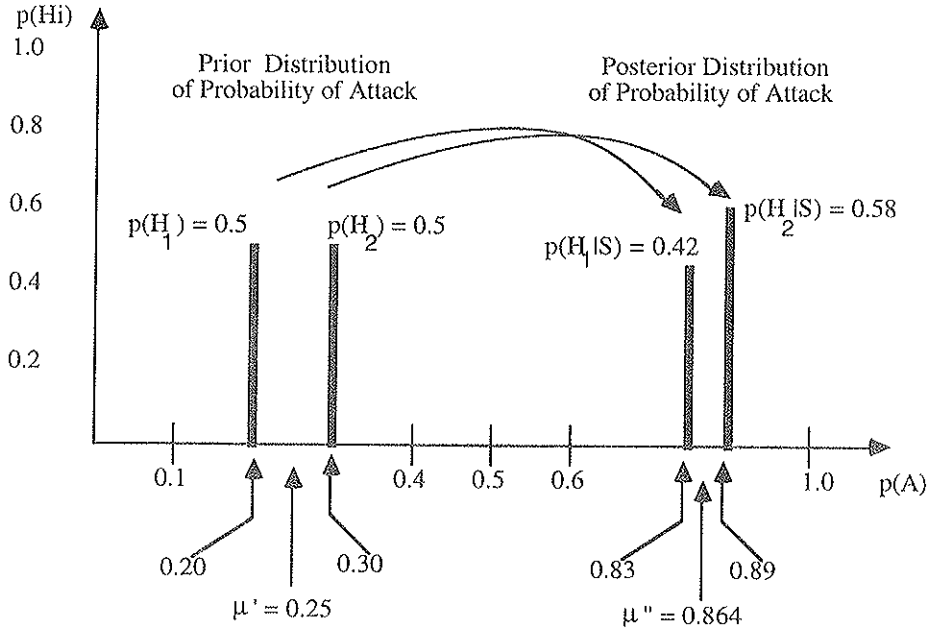


Fig. 2. Bayesian updating of the probability of an event based on new data assuming two discrete hypotheses H_1 and H_2 of prior probability 0.5, with prior probabilities of the event equal to 0.2 and 0.3 conditional on H_1 and H_2 respectively. *Source*: Paté-Cornell and Fischbeck, 1995.

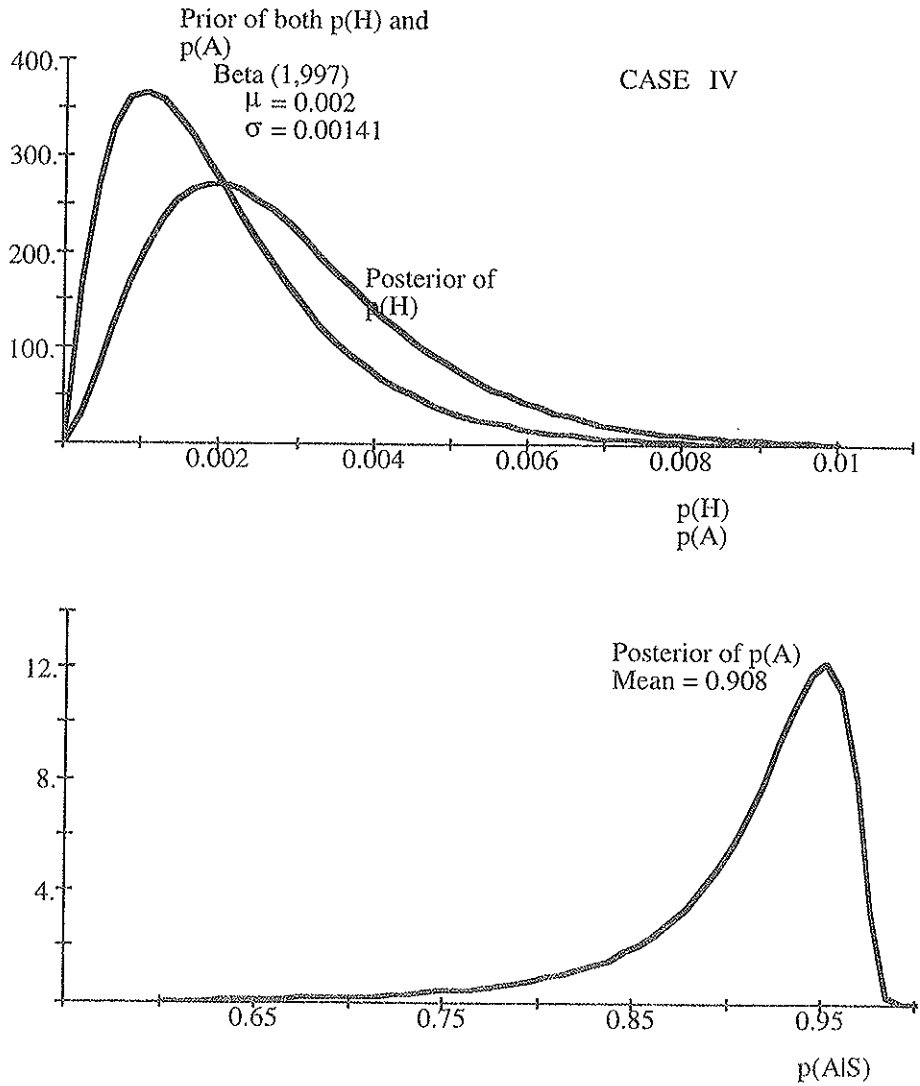
terior mean of 0.864 overall. Figure 2 thus displays the final risk uncertainties (here about the probability of the event of interest) in that discrete case based on two hypotheses.

Similarly, figure 3 shows the effect of the same Bayesian updating process for a continuum of hypotheses (e.g., the value of a parameter) and of probabilities of the event (one per hypothesis). Beta distributions are assumed for the priors, yielding a Beta distribution for the posterior probability of the event after updating based on the new piece of evidence (here, a positive signal). In this case, the probability of the event per time unit is assumed to have a prior mean of 0.002. The computation yields a posterior mean of 0.908 after updating based on the new data. What is important is to observe that even though the mean probability of the event is 0.908, the values of the posterior probabilities range roughly from 0.65 to 0.99 *with a non-uniform distribution*. This type of information is thus more complete than that provided either by a single probability estimate (the mean, here 0.908) or by a simple interval of results.[1]

This model can be used to assess the probability of a catastrophic ecological disaster, of a natural catastrophe, or of a serious disease when we do not perfectly understand the underlying phenomena and when the data can be misleading (the event may or may not occur given a positive signal).

The problem to which we applied this model (Paté-Cornell and Fischbeck, 1995) concerned the probability of a nuclear attack on the United States based on signals from the U.S. system of command and control which is subjected to both Type I errors (missed alerts) and Type II errors (false alerts). The decision problem, at the time of the study, was whether to respond to such a signal before a nuclear explosion on the U.S. soil, and how to account for the possibility of signal errors in the response decision. One problem was that this probability could vary according, for example, to fundamental assumptions made by experts regarding possible procedures and timing of attacks. It was assumed that the probability of an event then depended on which hypothesis was correct. We also considered the case where a continuous set of assumptions lead to a continuous distribution of event probabilities. The Bayesian reasoning described above allowed us to design a method to provide

[1] The equations that support this reasoning can be found in the original work where they are presented both for a discrete and a continuous set of assumptions (Paté-Cornell and Fischbeck, 1995). These equations describe a two-stage updating, first of the probability of the different hypotheses given a new piece of information of data, and second of the probability of the event or phenomenon given the signal. What is shown in this illustrative example is that the Bayesian reasoning allows representation of the posterior probabilities by a distribution representing uncertainty about the hypotheses, and not simply by a single point or an interval.



Note: Different scales used in each graph.

Fig. 3. Bayesian updating of the continuous probability of an event with a continuum of hypotheses. *Source*: Paté-Cornell and Fischbeck, 1995. The upper figure represents the updating of the probability distribution of the continuous hypotheses H. The lower figure represents the posterior probability of the event A.

the decision maker with a full representation of the state of information – including the uncertainties about the probability of an attack – after a reading of a positive signal from the command and control system.

The weakness of the technical system, in this illustration, is the possibility of errors in the different parts of the command and control system. The risk assessment problem thus starts with a logical analysis of the functions of the different components of this system to evaluate the probabilities of generating errors of both types in each of the elements (Paté-Cornell and Neu, 1985). This risk analysis permits linking the quality of the signal to the probability of attack accounting for the possibility of errors. Therefore, it also allows ranking the reinforcement alternatives (e.g., upgrading a specific critical radar) to improve the system in the long term. In the short term, i.e., in time of crisis, it allows setting priorities among the verification options immediately available.

Motivations for these probabilistic analyses

In each of these studies, there are reasons for an explicit treatment of uncertainties in analyses including both technical and human elements. Greed and ignorance are generally among them, for example under the form of production pressures without sufficient consideration for their actual effects. In the case of the tiles of the space shuttle, the technicians did not know the consequences of their short cuts, or for that matter, what there was under the orbiter skin at any of the spots on which they were working. Furthermore, to accelerate operations and meet contractual commitments, managers were setting schedule constraints without explicit consideration of the kinds of incentives that they were giving to the technicians, and of the effects of these incentives on the system's safety. Furthermore, NASA was under the illusion that everything was done perfectly, and therefore, that there was no need to set priorities in maintenance operations. It is the kind of overconfident ignorance and production pressures that can be the source of unpleasant surprises and prove fatal in flight.

In the case of anesthesia, there is also a desire on the part of hospital administrations to increase productivity in order to increase profitability, and on the part of the anesthesiologists, a willingness to yield to these pressures to increase their income. Furthermore, there is a definite reluctance among them to spend the time necessary for re-training, re-certification, supervision of residents etc. An explicit analysis of the potential effects of particular personnel problems on patient safety is an effective way to show the effect of greed on the risks to others' lives, and to challenge the excuse of ignorance ("we don't know if practitioner fatigue really threatens the patient").

In the case of multiple hypotheses regarding major risks, the problem is often that nations and industries can use the excuse of lack of certainty (or lack of agreement) to justify inaction and avoid necessary expenditures when partial information is available to support a sound policy decision. They can also overreact causing considerable economic damage based on little evidence when much more information is available. Risk analyses can thus provide some useful information to reduce the downsides of some basic human shortcomings. These include the tendency to avoid costs and efforts to protect others, to pretend that a problem does not exist when it is not certain, or on the contrary, that a catastrophe is certain to occur when it is still unclear or unlikely.

3. Human Shortcomings: Greed and Ignorance

Managing finite resources requires balancing consumption and investments as well as the mitigation of major risks. When we have only incomplete knowledge, such decisions present a challenge to human nature and its discomfort with complexities and uncertainties. Uncertainty can then become an excuse for selfishness and inaction, or on the contrary, for diverting large economic resources in one's interest, ostensibly to protect the public, but in fact against inflated risks and trumped-up fears. Scaring, confusing and baffling the public can satisfy large egos and specific economic interests. The conjunction of greed and ignorance can thus be a powerful and dangerous combination. But before seeking solutions, one needs to define, in this context, the meaning and implications of greed and ignorance.

Greed and limited resources

Human instinct is generally to prefer consumption today rather than tomorrow. Yet, saving today may allow more consumption later if investments generate wealth in between. Therefore, in economic terms, willingness to save is characterized by an equilibrium between wealth creation and propensity to consume as described by the marginal rates of transformation and substitution respectively (Samuelson, 1976). The former represents the rate at which society creates wealth from existing resources over time, and the latter, the rate at which people are willing to substitute consumption today for consumption later. In addition, two other factors should also affect fundamental choices but often don't given the scale of human life. They are the costs of irreversibility of the effects of some courses of action (recognizing, of course, that the world's evolution and life itself are irre-

versible), and the limits of the earth's resources which are themselves poorly known. The human propensity to accelerate consumption may lead to a refusal to spend now to protect other people, for instance against the hazards created by one's economic activities, in order to protect or increase the profits of these activities. But there is another side to that aspect of greed. It involves the diversion of public or private money in one's interest by cultivating a fear and inflating a threat for a variety of reasons. For example, the media may hope for more attention and sales, or a group of researchers may want to attract funding, or the competitors of an industry may want to force it to close for their own benefit.

Greed can thus be described in general economic terms as reluctance to save and to share resources, today or later, and to invest wealth for the protection of others and the benefits of future generations. It involves a general tendency to ignore the irreversibility of the damaging effects of one's actions, and the limitations of earth resources, including fresh water, clean air, and natural beauty. Greed can also cause the inflation and amplification of some threats to force the spending of public and private money in one's interests, or the dismissal of real threats to avoid forfeiting benefits.

Ignorance and the characterization of uncertainties

Ignorance, in the context of hazards, includes neglecting information that is already available, and sometimes a refusal to let facts guide reasoning and choices. More importantly perhaps, it is also lack of Platonic wisdom, that is, unwillingness or inability to recognize the limits of current knowledge. But the problem here is even more complex. To assess what we know *at any given time*, we need not only to recognize that these limitations exist, but also to characterize as precisely as possible our state of partial knowledge considering all available evidence. We need to do so because we cannot afford a wide margin of error in our estimation of the benefits of costly policies. Because we are not infinitely rich, the limited resources of the planet have to be wisely spent and shared with others, now and later. Because we do not have full knowledge of some potential threats to mankind (and yet cannot always wait for that information before acting), we have to make choices under uncertainties to best use limited resources. This is why we need quantitative methods that are designed to assess the magnitude of existing or anticipated risks and the potential benefits of risk reduction measures (Lave, 1987b). These methods, however, are still rather coarse and need to be improved. These improvements will come both from increases in computational power and from reasoning and algorithms more sophisticated or more robust than those that are in use today.

4. Management of Risks and Tradeoffs

Probabilistic risk analysis and the relevance of risk magnitudes

Risk is defined here as the uncertain losses that can result from the existence of a hazard (e.g., hurricanes). The quantitative analysis of a risk permits us to characterize it by the probability and the magnitude of the consequences of all scenarios that can be envisioned (Kaplan and Garrick, 1981; Apostolakis, 1990). These scenarios can involve accident sequences, for example, a sequence of events leading to a catastrophic failure in an industrial plant. They can also be conjunctions of hypotheses, events, and/or random variable realizations leading to an outcome of particular severity such as a specified increase in the earth's temperature. A well-conceived risk analysis is based on the identification of an exhaustive and mutually exclusive set of such scenarios. Therefore, the first step in this analysis is generally a logical analysis of a system's functions to identify its failure modes, i.e., the conjunctions of events leading to a system failure. For a given time interval, the results can include the probability of a dangerous event (or the probability distribution of its future frequency), the probability distribution of the outcomes or losses associated with a hazard, or a measure of the individual risk. The latter is generally described as the probability of death of an individual exposed to the considered hazard. The results can also be reduced to the expected value of the overall losses for the same time interval. The risk magnitude can thus be represented by different scalars that allow risk rankings and comparisons.

Clearly, risk magnitude is only one of the factors that should guide risk management decisions (Fischhoff *et al.*, 1981). Distribution effects (who enjoys the benefits and who is subjected to the hazard), and the facts that some risks are uncontrollable, involuntary, new or unknown, must also be given attention (Slovic, 1987). Therefore, the strict order of risk magnitudes does not – and should not – rule priorities in risk management to the exclusion of all other factors. Yet, quantification may help focus the debate on the relative importance of different risks and on the contribution of different factors to a particular risk.

This is especially important when misperceptions of threats and priorities are shaped by unfounded fears (Slovic *et al.*, 1982) or by the opinions of experts with strong positions at stake. Fear is a great motivator and an essential safeguard of mankind, as well as one of its weaknesses. As such, it can be an effective way for the media, political powers, or others to frighten the public to serve their own interests. Fear is also painful, and unpleasant realities – especially if they affect others – can sometimes be ignored.

Indeed, burying one's head in the sand is a common defense mechanism against fear of the unknown and unwelcome responsibilities. The great threats that weigh over the world are not necessarily those that make the largest headlines. Therefore, quantifying the different risks and debating the results may help clarify the issues, and either deflate an overblown threat or bring to light an underestimated hazard.

Tradeoffs

Mankind continually faces tradeoffs between prudence and the desire for innovation and improvement. The first human being to eat cherries risked painful death by poisoning. Heating homes has perhaps been one of the greatest improvements of comfort and health, but it has generated deadly fires, fumes from burnt wood, smog from coal burning, and more recently, the damage of acid rains. Few existing hazards do not involve such tradeoffs between risks and benefits. Global warming may be the cost of cheap energy, severe pollution that of industrial wealth, and devastating natural catastrophes the price of living in attractive areas prone to floods and earthquakes. Similarly, the risk of catastrophic economic crashes is a downside of the newly found riches created by the globalization of world markets. The choices associated with wars, ethnic conflicts, religious persecutions and terrorism are still more complex; but for the people involved at the time, they present similar tradeoffs given the perceived benefits of defense – or expansion – of their land, their heritage, their religion, and what they see as their interests.

The observation that hazardous activities (or voluntary exposure to a hazard) simply represent acceptance and resolution of tradeoffs deserves a few caveats. First, of course, many risks such as global warming are involuntary and uncontrollable to most individuals, even though mankind as a whole creates the problem. Second, it may not be the same people who incur the costs and enjoy the benefits of an activity; for example, those who live downstream of a dam are exposed to the risks of its failure whereas a much larger population enjoys the benefits of the electricity that it generates. Therefore, the effects of risk-benefit tradeoffs are not only unevenly perceived but also unevenly felt. Third, if only it were feasible, many reasonable people would gladly eliminate some hazards such as the use of hand guns by simply banning the product or the activity. In the absence of full information, "prudent avoidance" seems wise (Morgan, 1992) unless it ends up stifling potentially beneficial efforts and generating costs without reasonable hopes of benefits.

Judgments about risk taking

Unfortunately, "sensible" universal criteria of risk acceptability have no firm foundations: risk attitudes vary widely among individuals, among social groups, and over historical times. There is no universal point where risk aversion is clearly absurd, where risk taking becomes "gambling", where courage becomes temerity, or where prudence becomes timidity. Instead, often in addition – and in contradiction – to a "zero risk" statement, what generally exists is an ill-defined, temporary and tacit consensus about acceptability of non-zero risks. In the end, however, it is seldom the decision itself that is judged in practice, but the result or the outcome of an action without sufficient consideration for randomness. A sound decision may have been followed by an unlikely failure, a stupid choice may have been rewarded by chance, and in every day life, that is often what supports judgment. Yet, in order for the experience to be of value in the future, the norm should be applied to the decision itself and to the risk *a priori* rather than the outcome.

The objectives of risk analysis

Under such circumstances, social incentives and prudence may encourage taking minimal risks, especially in public policy decisions. Yet, "the worst is not always certain" and little may be known at the time of a decision about the actual risks and benefits of many human endeavors. Managing large-scale tradeoffs does not imply freezing the world in its current state because the opportunity costs – i.e., the foregone benefits of alternative investments – may be unbearable, or because it may not be feasible: nature may have other plans. On the other hand, industrial as well as private activities involve a spectrum of risks, some unnecessary and some unavoidable. Unless the activity is forbidden, a number of risk mitigation measures may be available given the current state of knowledge.

Therefore, risk analysis has several objectives. The first one is to assess the foundations of fears so that fears alone are not allowed to drive world policies or individual decisions, but that real threats are addressed in time. The second one is to guide the choices of investments in basic research towards domains and topics that will provide the best support for risk mitigation policies. The third objective – and perhaps, the most important – is to weigh the costs and the benefits of various risk management policy options and to set priorities among them.

Information as a basis for decision and action

To resolve the tension between mankind's yearning for a better fate and its desire for protection and continuity, one may simply rely on the natural

dynamics of human groups. Unfortunately, decisions can then be made regardless of facts or fairness by those individuals who are the most determined, and/or control most of the resources. Greed and ignorance may then constitute the only motivators and the greatest obstacles to a prudent yet innovative management of the earth's resources. Democracies, in principle, feature processes that can balance the interests of individuals and attenuate the effects of individual greed, but often within national borders and in the present time. Even though they generally involve freedom of information, democracies are also poorly equipped against overconfidence in scientific knowledge because they have difficulties handling uncertainties. People prefer certitudes, and for fear of popular backlash, politicians – and sometimes scientists – are often unwilling to admit that in many activities, there is a residual risk whose total elimination would require resources well in excess of the expected benefits. In spite of an evolution towards realism about risks (in particular in U.S. laws), the recognition and the treatment of uncertainties remain weak features of many decision processes, both in industries and in governments (Paté-Cornell, 1996b).

Given the scale of the consequences of some major threats, the complexity of risk management choices, and the limitations of resources and knowledge, a substantial investment in science and engineering research seems appropriate. The questions are: where and how should this investment be made, what information and technologies could be most beneficial, and how should residual uncertainties be treated in policy decisions?

5. The Role of Science and Engineering

Science, engineering, medicine and economics, all have a role to play in the management of major risks. Science formulates, explores and tests new hypotheses about fundamental phenomena and tries to discover new mechanisms of defense and of protection against catastrophes. Once the problems are understood, engineering and medicine can develop and implement mitigation techniques. Engineering and economics combined can also provide analytical tools to support public policy decisions by assessing the value of prospective scientific information and of various risk mitigation options. Yet, both science and engineering can also contribute to the creation of risks and hazards. Their proper roles could thus be summarized as first do no harm, then solve existing and possible future problems.

First, do no harm

The human creation of a risk can often be viewed in the context of a trade-off of the type discussed above. To avoid some of those dilemmas,

one can argue that science should first avoid the creation of "hazardous" knowledge that opens possibilities of action. Even if that knowledge exists, engineering could then be prevented from transforming it into devices or procedures that could have dire implications for mankind. The problem, of course, is that at the time of scientific discovery and technology development, there is often a strong perception – at least by some – that the benefits outweigh the costs.

Faced for thousands of years with a similar dilemma, medicine leaves the judgment to the individual conscience. Based on the experience of the Hippocratic oath, Serres for instance, proposes a voluntary oath by which scientists would vow (among other things) to make sure that the product of their work would not promote "violence, destruction, death, enslavement or inequality" (Serres, 1997). This is certainly a noble goal but one cannot always anticipate what will be harmful in the future. Also, in practice, these criteria can be difficult to reconcile with the very societal needs that the scientists' work might serve in the short term. The development of new techniques generally brings real benefits to their authors, to those who fund them, and often to their nation as well, thus increasing inequalities at that time but generally enriching mankind in the long run. If research and development must be restrained, a distinction can be made between science that generates knowledge and engineering and medicine that put it to use, and hence find themselves at the interface between science and society. One might argue that whereas restricting acquisition of knowledge is both infeasible and counterproductive in the long run, safeguards could be placed around engineering if it puts that knowledge to questionable use. In reality, the boundaries between science and engineering are often unclear, and once a phenomenon is understood, its applications are often already available in practice.

In spite of these difficulties, it seems appropriate, before addressing the mitigation of man-made or natural risks, to worry about the social hazards that science and engineering themselves might generate. Short of strong international agreements, legal safeguards are limited by national boundaries and may simply send the activity in question offshore. It may thus be wise – if not fully effective – to put that choice explicitly in the hands of those who actually decide whether and how to work on specific topics, as is the case, for instance, in the codes of professional ethics of some engineering societies in the U.S.

Science and engineering resources for risk mitigation

Whether or not they have created the hazard in the first place, science and engineering have an important role to play in the mitigation of existing

risks. Faced with a spectrum of threats, societies as well as individuals have a number of possible choices. They can ignore the problem and decide that they will later adapt to the circumstances. They can ban an activity, a local occupancy, or a type of product. They can identify signals and provide warnings of danger. They can also provide financial insurance. More effectively, they can actually try to mitigate the risk by eliminating or attenuating the hazard at its source, or by reducing the vulnerability of the people or the systems exposed.

Consider, for example, the case of seismic risk. Science may be able, one day, to understand the fundamental earthquake mechanisms. In the best of all worlds, it may then be possible to prevent the locking of faults and the accumulation of strain energy at the interface of the tectonic plates. Short of such feats, precursor signals might be observed in specified areas and permit taking protection measures (Keilis-Borok, 1972, 1980). More realistically, the future frequencies of earthquakes of given magnitudes might be established better than they are now. Engineers can then provide techniques of reinforcement of structures adapted to the local seismicity. Public policies then have to be designed to implement building codes that strike a balance between costs and residual risks. Finally, governments (or perhaps private entities) also have to decide how and when to provide earthquake warnings that are both "safe and effective", protecting people without exposing them to disruptive false alerts. The best policies will generally involve a mix of earthquake engineering and prediction, again, within the limits of available resources given the spectrum of existing threats. Under such circumstances, priorities have to be set and economics matter. If one agrees that protecting more people rather than less for a given cost is an appropriate objective, it is important to know the magnitudes of existing risks to estimate the benefits of mitigation options.

The engineering risk analysis methods described earlier provide the logic to treat the available data, and the techniques to quantify the current degree of knowledge through probability. One of many practical problems is the analytic instability of results that can be caused by complexities (multiple feedback mechanisms), discontinuities and non-linearities. Systems analysis generally permits identification of zones of relative stability and of zones of parameter values where either the existing models or their inputs are so unstable as to make the results meaningless. In that case, one can try to restructure the problem, or to identify the domains of input values where the results are unstable and if possible, to provide an envelope of the results. Increased computational power can then be of some help, but only to the extent that the models are correct.

Value of information

New information is costly and scientific research priorities can be set in many different ways. Curiosity has long been the prime motivator of research, although the downstream needs have generally played a role as well. In the context of decisions under uncertainty, the search for useful knowledge can be focused more sharply than by instinctive curiosity alone. The engineering risk analysis techniques that permit risk estimations also permit computing the (decision) value of prospective information based on two factors (Howard, 1984). The first one is the probability of getting different results in the course of the considered research, and the second is the effect of these different results (e.g., the reduction of uncertainties) on the desirability of a particular course of action.

Knowing more before making a critical decision permits avoiding costly mistakes. This improvement of the anticipated decision *outcome* at the *onset* of the research is what is called its value of the information. It should therefore be noted that this concept of *value*, which is applied to the prospective gathering of new information, is restricted to the context of a particular decision. It does not cover the beauty of a totally useless but marvelous discovery, which obviously has value to mankind in a different way. The notion of value of information is a useful way however, to assess *a priori* the practical benefits of acquiring different types of knowledge – including from basic research – based on their prospective effects on specific policy decisions. An estimation of the value of information then allows allocation of a budget of research about a particular risk among different research groups given their skills and focus of interest, and among different topics given their importance in the context of policy decisions.

Clearly, to provide the necessary information, a risk analysis should be adapted to the needs of the decision process in which it will be used. But the magnitude of the risk results should not be influenced by the decision to be made and should remain an independent input of that process. The risk analyst who conducts these studies, structures the problems, gathers the data and computes the results must carefully separate facts and preferences in the gathering of risk information. He or she must ensure in particular, that the experts do not inject their own values in the assessment of the facts in order to influence a decision. As a classical illustration, a well-performed seismic hazard analysis for a given site should provide the same result regardless of whether one considers building there a nuclear power plant or a chicken coop. But of course, for a nuclear power plant, the level of resources and efforts invested in the uncertainty analysis should be higher to provide greater confidence in the results and permit prudent decisions.

6. The Nature and Characterization of Risk Uncertainties

Uncertainties and sources of data: synthesis of available information

The analysis and the quantification of risks, as currently performed, are based on mathematical statistics when feasible and, for poorly known phenomena, on systems analysis and probability. Statistical analyses are possible – and useful – to describe the effects of a hazard that has been previously observed and for which abundant (and well controlled) statistics are available. Unfortunately, that is not often the case. In the absence of overall statistical data, one must perform a synthesis of the different aspects of a problem, based on various sources of information, bodies of evidence, and even several fields of science, in order to characterize potential risks (see, for example, USNRC, 1975; Garrick, 1984). For example, seismic hazard analysis can seldom be supported by sufficient statistical data because the timeframe of human experience and the time span of other retrievable data are too short compared to relevant geological time scales. Furthermore, seismic *risk* also includes the robustness of different kinds of structures under seismic loads, and this capacity is often poorly known, even for new well-engineered types of buildings. In such cases, one needs to decompose the problem and to gather different types of information from all legitimate sources about both seismicity and building performance (Cornell, 1968, 1978).

Data necessary for risk analyses are generally of three types: statistics whenever available, physical and engineering models, and expert opinions. The use of statistical evidence to generate probabilities often requires an understanding that many phenomena do not occur in a steady state, and that existing statistics may have to be updated to actually represent the events of interest. Physical and mathematical models provide schematic representations that are always based on some level of approximation. Finally, expert opinions are necessary to characterize epistemic uncertainties, which reflect incomplete knowledge of basic phenomena.

This subjective element of risk analysis often troubles decision makers (Ellsberg, 1961), but if the hazard is poorly known or the environment has changed, it may be the best information upon which they can rely. The alternative to this subjective but carefully reasoned, treatment of information may be a mere guess of the risk magnitude, reflecting a much higher level of aggregation, that is, guessing the final results as opposed to addressing separately the better known components of the problem. If this problem is properly structured, modeling and deductive reasoning can reduce the uncertainties by treating its components at a lower, more manageable level of analysis and gathering the best information for these different parts.

In particular, it allows calling on the most qualified experts in each relevant domain, for example, seismologists to assess earthquake hazards and engineers to estimate the seismic capacities of buildings.

These probabilistic analyses are obviously not substitutes for full scientific knowledge, but engineering tools that allow the best representation of a state of partial information about a particular problem at a given time, in order to provide decision support. These tools are limited by the logical impossibility of accounting for new, unknown, chaotic, and unpredictable phenomena. They also require data that are sometimes difficult to gather. But their greatest limitation perhaps, is that these methods are complex and can be badly used. Common errors include, for instance, truncating or poorly formulating the problem, failing to account consistently for the different uncertainties, or assuming variable independence and model linearity where they do not exist. Intentionally or no, these techniques can thus be misused to deny the existence of real hazards or to exaggerate the effect of minor ones in order to influence a specific decision. Yet, if properly applied, they are useful, first because they provide a language of communication among the experts, the public and the policy makers, and second, because they permit a synthesis of available knowledge about the effects of different risk mitigation options.

Uncertainties and policy decisions: example of global climate change

Society often faces the need to make decisions before full information is available.[2] It was true when electricity was adopted in spite of claims from some of the brightest minds at that time that electricity was going to have a number of dreadful effects. It is true now as we face the possibility of a global climate change – over and above natural climate evolutions – induced by human activities and the accumulation of "green house gases" such as carbon dioxide in the atmosphere. On the one hand, delaying action can trigger irreversible damage and catastrophic changes. On the other hand, some options may be ineffective because they have large economic consequences in themselves and the problem may be elsewhere.

The models used in the assessment of the effects of global climate change are based on limited knowledge of the fundamental phenomena, for instance, the role of the clouds and of the oceans (IPCC, 1996). Although a general consensus seems to exist among the scientists involved, the very existence of this consensus does not constitute proof that it represents real-

[2] Section 5 of this paper is based in part on an editorial essay published by the author in *Climatic Change* in 1996. Reproductions were made with the consent of the journal editor.

ity. Uncertainties remain and many of the generally accepted assumptions need to be revisited. The effects of *some* of these uncertainties are represented in the published results (e.g., temperature increases) under the form of confidence intervals. Yet, all equal parts of these intervals do not have the same probability, and many segments outside of these intervals do not have a zero probability. Much work is needed to improve the information used for public policy in response to a threat of global climate change. Some of it is fundamental research, and some of it involves a better representation of the information that already exists.

When science can progress quietly, independently from the pressures of public policy making, the scientific community has ample time to wage its internal battles and to prove or disprove each element of the problem. There is no need to synthesize the state of knowledge until the problem is considered resolved by most. In that context, errors may not matter much. The speed of light, for example, was measured over many years with different levels of accuracy (Henrion and Fischhoff, 1986) until available instruments and methods allowed general confidence in the results. When decisions need to be made along the way, based on partial and incomplete information for private purposes or public sector regulations, one does not have the luxury of taking the time to reach a complete, unquestioned consensus. In that case, as discussed earlier, the available information, imperfect as it is, must be synthesized at a particular stage to represent as closely as possible the state of knowledge at that time.

Bayesian reasoning

This synthesis must be made under uncertainties. In that case, the best approach – although not an easy one – is to combine existing information into Bayesian probabilities (e.g., Savage, 1954; de Finetti, 1974) that represent a collective "degree of belief" in the different hypotheses given the evidence. This Bayesian reasoning is based on two fundamental principles of logic. The first one is that the probability of a given hypothesis AND the evidence observed is equal to the prior probability of this hypothesis multiplied by the probability of observing the evidence given the hypothesis (Bayes theorem). The second principle is that the probability that a piece of evidence is observed is equal to the sum, for all possible (mutually exclusive and collectively exhaustive) hypotheses, of the joint probabilities of this piece of evidence AND of each of the hypotheses (total probability theorem). These two fundamental equations (and their derivatives) permit computing the probabilities of the different hypotheses given the evidence that has been observed. Afterwards, a well structured model formed of several "sub-models" – e.g., of

gas emission, trapping and reflection of heat, feedback mechanisms, etc. – may allow computation of the probability distribution of the outcomes for example, the global increase of temperature for a doubling of the carbon dioxide concentration. The modeling of uncertainties is based (1) on the probabilities of the different hypotheses for each of part of the overall model (oceans, atmosphere, etc.), (2), on the distribution of parameter values for each hypothesis and each sub-model, and (3), on the computation of the consequences for each combination of sub-model, hypothesis and parameter values. In this computation, it is essential to use conditional probabilities to correctly represent dependencies among sub-models, hypotheses and parameter variables and, of course, to eliminate impossible combinations.

The first problem in this exercise is to identify a set of possible hypotheses that can be envisioned for each sub-model given the state of information. Second, the procedure requires a strict structuring of this set of hypotheses (they must be mutually exclusive and collectively exhaustive). Third, one must encode from experts the probabilities of these hypotheses *a priori*, and the probabilities of the evidence (i.e., the statistical data) given each hypothesis. Fourth, these opinions must be aggregated to obtain probabilities (marginal and conditional) that represent the collective wisdom of a group of experts that often disagree. In the end, the results of the analysis must clearly reflect all uncertainties involved. This can be done through a probability distribution for the probability of an event of interest, through families of risk curves that represent distributions of the future frequencies of exceeding different consequence levels (e.g., losses), or through intervals of risk magnitudes (e.g., future temperature increases). The illustration presented earlier in Section 2 was an example of this kind of probabilistic reasoning.

Expert opinions and disagreements

Given that they have had different experiences and that they start from different beliefs, the experts generally have different prior probabilities of the hypotheses and different conditional probabilities of the evidence given each hypothesis. When the probabilities from the different experts have been encoded, they must be aggregated in order to obtain a global estimate. This aggregation can be approached from a mathematical point of view, assuming the existence of a "super expert" who among other things, accounts for dependencies among experts (e.g., Winkler, 1986; Morris, 1977). It can also be treated iteratively without any interaction among the experts, e.g., by the Delphi technique (Dalkey, 1967). In reality, the most promising approach seems to be an interactive procedure in which the experts exchange and debate all available information.

One such process has been recently developed in the domain of seismic hazard assessment (Budnitz *et al.*, 1998). This method has been applied to other related fields such as the risks of volcanic eruptions, and the general idea seems promising. One essential characteristic of this aggregation procedure is that it must be collaborative in nature and that its success is essentially incompatible with the adversarial process that prevails in the U.S. court rooms. The interactive procedure mentioned above involves experts who act as proponents of their favorite theory, and experts (sometimes the same ones) who act as evaluators of each alternative given the evidence. Finally, the experts are asked to act as integrators of whole probability distributions. This process thus requires that the experts be willing to consider the problem from different angles in order to come up together, not necessarily with the "truth", but with probabilities that are as close as possible to an accurate representation of the common current state of knowledge *at that time*. Yet, in the U.S. scientific community, there is currently a real danger that the adversarial process could become the norm in this particular procedure. This is true because as risk assessment may become part of the law, any judicial review is going to push the process in that direction.

The pros and cons of an adversarial process

There is nothing wrong with the adversarial process in the long-term search for scientific truth. Indeed, it is the scientific community's norm by which theories emerge, are debated, tested and opposed to other theories over time. In the procedure described above, the challenge and defense of a contending model may be a good way to get to the bottom of the data base in the "proponent" phase. After a debate on the quality of the available information, however, the adversarial process must stop in the technical integration phase where one needs to gather all pieces of an incomplete puzzle and to form a collective judgment based on all available information. At that point, the objective is to synthesize the state of the art, recognizing and weighting the possibility of different assumptions. The adversarial process can then get in the way of a rigorous integration of each piece of evidence in probabilistic estimates because it provides incentives to all parties involved to *truncate* the information in order to support their favorite hypothesis. Therefore, it can result in a polarization (and a weakening) of the evidence base that obscures the debate, sometimes giving too much attention to the extremes, with no attempt to seriously consider data that may reconcile the different theories. Under those conditions, the adversarial process can lead to anything. The result of this tug-of-war has no reason to fall anywhere near "the truth", that is, an accurate description of knowl-

edge at the time. It is therefore all the more important to develop and to use methods of aggregation of expert opinions that explicitly attempt to escape the default solution of a confrontation in which the evidence is simply truncated by both sides to fit extreme hypotheses. It is also important to recognize that the sheer number of experts on a given side is not a measure of the probability of a hypothesis but may, for instance, be biased by the mechanisms of research funding.

The social process of combining expert opinions

Obviously, the process of encoding and aggrégation of expert opinions is a social as well as a scientific process. The social aspects involve the choice of experts and their mode of interaction; the scientific side involves the logical treatment of probability in risk analysis. It is essential that both be handled well. Whereas the logical mechanisms are relatively straightforward, the social quality of the process is more difficult to assess. First, the choice of experts has to be broad enough to permit representation of all sides and viewpoints. Again, there is no reason to believe that the probability of a hypothesis is equal to the proportion of experts who believe in it. Indeed, there may be a trend in research that has thinned the ranks of the opponents of a popular theory. This does not necessarily mean that the popular view is incorrect, but perhaps that this hypothesis and its implications are politically more palatable to many. To be sure, it may be more appealing because it is correct and the others are not, but this is in no way guaranteed.

The social process of collecting and aggregating expert opinions must thus require that all options be considered, and that a theory be rejected only if it contradicts the basic laws of physics or the evidence. The experts must therefore be willing to consider the possibility of mechanisms that are different from their pet theories and to assess their probabilities in the light of the complete evidence. In particular, an expert who puts a probability of one on his model and zero on others' should simply be excused from this exercise.

Whereas randomness (or aleatory uncertainty) can be treated through standard statistical procedures, incomplete knowledge about fundamental phenomena (or epistemic uncertainty) has to be treated through subjective probability. One could simply choose to encode the information from each expert and to leave the aggregation task to the decision maker who then has to weigh explicitly each of the models assessed by the experts. This process may be arbitrary and led by political concerns rather than by the scientific evidence base. By default, for instance, it is often the simple number of experts of a particular view that eventually plays the role of the 'weights' attributed to the possible assumptions. An alternative is to simply present

separately the spectrum of final results (including both hypotheses and parameter values) provided by each of the experts (Morgan *et al.*, 1995). If all experts are required to provide not only a central value but also a confidence interval, the tendency of the decision maker may then be to look for a point or a segment that would be common to all confidence intervals. Experts, however, tend to *underestimate* uncertainties, and it could well be that the more informed and qualified the expert, the larger his or her uncertainty band. It is thus possible that the common interval is simply driven by the judgment of the most stubborn (and not necessarily the best informed) who describes his knowledge by the narrowest interval. A probabilistic treatment, even though the opinions of experts are only that, permits a logical treatment of the inputs.

Clearly, in all cases, the composition of the group of experts and the classic biases in expert opinion elicitation (Tverski and Khaneman, 1974) affect the results. Therefore, the methods that are most likely to provide a reasonable degree of objectivity are those that focus on the gathering of a group of well informed and socially adjusted individuals, on the construction of a complete set of hypotheses, and on the assessment of axiomatically correct probability distributions based on all scientific evidence. At that stage, an adversarial procedure can only be counterproductive if it leads to the truncation of the evidence base to focus on extremes. In the court room where the objective is a verdict "beyond reasonable doubt", it may be appropriate. In a public policy arena where scarce resources must be allocated and therefore, priorities must be set, an adversarial process is likely to lead to bad economic decisions and eventually to failures of public policies.

7. Limitations of the Probabilistic Method

In spite of the insights that they provide, the methods of risk analysis are limited in many dimensions. Some of these limitations are fundamental in nature and reflect weaknesses of these methods. Others can be, in principle, more easily addressed and remedied because they reflect mistakes in applications of these techniques. Only a few examples are listed here; this list may be far from exhaustive.

Subjectivity and falsification

One of the ultimate tests of scientific reasoning or method is whether the result can be proven false (Popper, 1962). In probabilistic analysis, on the one hand, errors of logic can be pointed out; so, in that sense, probabilities can be

"falsified". On the other hand, reasonable experts can interpret differently the implications of the same evidence for the probability of an event. This may happen because they have estimated differently (1) the prior probabilities of the same event (or hypothesis), or (2) the probabilities of obtaining the observed data given the event or the hypothesis. Those elements of subjective reasoning, which are often implicitly at the heart of the debate among experts, are some of the most difficult disagreements to resolve.

Completeness

It is often impossible to ensure that the set of hypotheses considered at any given time is complete, i.e., that the available evidence could not possibly support other possibilities and scenarios. For example, it could be that the increase of temperature that has been observed in the last century is caused entirely by human activities. But although less likely, other hypotheses are plausible, including different conjunctions of natural and man-made effects. Furthermore, the possible hypotheses are often difficult to structure into a collectively exhaustive and mutually exclusive set, which is one of the requirements of probabilistic analysis. This means, for example, that a number of factors may influence the climate simultaneously, but that we have not yet identified them and separated their effects. This lack of assurance of completeness is not in itself a weakness of the risk analysis method, but it is a fact of life – whether or not a quantitative analysis is performed – that can affect the results of risk analyses.

Insufficient data and problem structuring

Enough data may not be available to assess probabilities in a way that satisfies the decision makers. Even though it is understood that the method allows processing small amounts of evidence to update prior estimates, the results may not be "firm" enough for many people to act on that basis. What is the alternative? When that is the case in practice, decision makers often prefer to rely on their instinct rather than on a more sophisticated analysis based on what they consider "soft" information. In other instances (for example, to assess the health effects of 60-Hz electromagnetic fields on human health), the basic problem is the difficulty to structure the analysis so as to be able to make use of available information. The risk analyst then faces the task to formulate the question in such a way that the experts are able to provide probabilistic assessments that reflect their experiences. This problem formulation is sometimes extremely difficult, and there does not seem to be a general solution to that question: structuring an analysis is an art as much as a science.

Vulnerability of results and opportunities for manipulations

The results are generally sensitive to stated or unstated assumptions and can be manipulated by interested (and perhaps, unscrupulous) parties. One critical issue is the framing of the problem and the sensitivity of the results to these boundaries which, as mentioned earlier, can sometimes be set so as to obtain the desired results. For example, some risk analyses of alternative means of generating energy have focused exclusively on the production process, leaving aside problems of fuel extraction and transport as well as the issue of waste storage. In other instances, the redistribution effects are ignored; some public policies can simply displace the risk from one group to another. Again, the problem here lies with the use of the tool, not with the nature of the technique itself.

Unwarranted implications of risk acceptability

Correct results can be misused and misrepresented to suggest risk acceptability. For instance, the way some results are displayed ("exposure to risk X is equivalent to smoking n cigarettes per day") seems to imply that in order to satisfy some criterion of "rationality", risk magnitudes (for example, individual risk) alone should determine risk acceptability. Yet, comparing for example, industrial risks to risks of cigarette smoking is misleading because it implies that the acceptance of a *voluntary* risk of given magnitude should rationally dictate the acceptance of an *involuntary risk* of the same magnitude, possibly by different people. Therefore, what may be seen as a simple and effective way to communicate risk magnitudes through every day life experience, may be wrongly interpreted, for example, as implying that because some people smoke, the general population should accept all risks of comparable magnitude.

Treatment of human errors

Finally, this kind of analysis generally does not treat human errors, human actions, and human adaptability as well as it treats physical and technical variables (Perrow, 1984). This is probably the case because there is widespread belief that human behaviors are more difficult to predict and more variable across individuals. Yet, this may not be true and some models, such as the one described earlier that relate physical variables to human decisions, actions and management can be useful for that purpose (e.g., Paté-Cornell, 1990; Davoudian *et al.*, 1994).

This list of course, is not exhaustive. Assumptions – conservative or not –

may be hidden in the analysis. Uncertainties can be treated in an inconsistent way so that the spread of the results is underestimated. Chaotic behaviors (Gleick, 1988) may exceed the power of probabilistic methods and system complexity may require other approaches (Nicolis and Priogine, 1989; Gell-Mann, 1994). Specifically, the models linking the different variables may present such non-linearities that the results are extremely sensitive to the inputs and therefore, difficult to interpret and to use. In that case, one option is to try to restructure the model to define classes of scenarios that are more stable. Another option is to try to identify input domains in which the results are stable and unstable, and to find an envelope of the outputs where instabilities occur. In other words, one does what one can.

In the end, the question is whether other methods of reasoning, when little is known, are better, for the purpose of allocation of scarce resources for the protection of people and the planet, than this type of probabilistic reasoning.

8. Some Ethical and Moral Issues

When performing and using risk analysis, one makes several fundamental assumptions, for example, that the evidence matters, that uncertainties can be quantified, and that one can separate facts and preferences. More importantly, it is assumed that if a decision has to be made before it is too late, it is acceptable to use analytical results that do not pretend to represent the ultimate "truth" but a snapshot of what the evidence supports *at that time*. This snapshot is necessary to allocate existing resources. The first issue is thus the moral implication of economic constraints in risk management and of tradeoffs between safety and consumption.

The role of economics

A critical moral question is whether economics matter at all in the domain of risk mitigation. One can argue that above a certain threshold of individual risk, no one should be subjected by others to involuntary risks regardless of economic considerations, and that below that threshold, cost-benefit analysis is justified because more lives could be protected by investing resources elsewhere (Paté-Cornell, 1994). Where this threshold of tolerability resides remains an issue and varies with the circumstances. An alternative to a quantified "acceptable risk" is to design an acceptable decision process in which all affected parties are involved. The problem, however, is that all Pareto-optimal deals – in which all parties believe that they are better off if the deal is made – may not be ethically acceptable. For exam-

ple, a poor community may find it to its advantage to accept to store, for a price, the wastes generated by a richer one, but the potential exposure of its inhabitants to toxic risks may raise issues of environmental equity. In general, however, because no one is infinitely rich, economic considerations are unavoidable. Nations, communities and individuals must make choices that involve investments, consumption and risks, recognizing that the consumption level in itself is also a factor of longevity.

Discounting

A corollary to the relevance of economics is whether it is ethical to discount the future effects of hazards and risk reduction policies. Discounting implies that future risk reduction benefits are smaller in present value than the same benefits today, and therefore that discounting may lead to policies and choices that seem selfish at first sight. Yet, as discussed earlier, discounting future cash flows reflects the fact that society between now and then generates wealth and knowledge. Therefore, discounting costs and *not* the risk reduction benefits would always make it look attractive to delay spending in order to reap larger safety benefits later given the present value of the costs involved. Doing so, however, simply reflects an error of logic and would systematically shortchange present generations.

One simple principle that can guide the reasoning in this domain is that what is judged desirable to us today should also be judged desirable as part of our legacy to future generations (1983b). In the domain of individual risks (probability of individual death) where cost-benefit analysis is not acceptable because the risks are intolerable, one can adopt the rule that what is intolerable today should not be considered tolerable for the future generations. Therefore, under those circumstances, the costs of eliminating a hazard for the future generations should conservatively be accepted by the current ones, even though there is a chance that between now and then, a solution might be found that would save some expenditures today. In the risk domain where economics matter today in the sense that the costs have to be justified by the benefits, the same reasoning should be apply to future generations. To the degree that discounting reflects the creation of wealth between now and then and therefore, that our descendants at the considered time will be richer than we are for it, the costs and the benefits of risk mitigation should be both discounted. In this assessment, irreversibility can be accounted through a finite (possibly very large) cost. Then, discounting at a rate that reflects this expected increase of wealth simply implies that the same amount of resources should be available for individual risk mitigation at future times as is available now for the present generation.

Knowledge as power

The probabilistic method itself assumes that experts have a knowledge that is unavailable to the lay people and this knowledge, which is an important input to the policy making process, is in effect a legitimate source of power. In reality, experts can be biased in the sense that they have overconfidence in a particular theory. They can simply ignore hypotheses that they do not like for whatever reason, and they can use their position of expertise to promote an agenda based on their preferences (as opposed to what they know). The problem is that in the probabilistic world of subjective treatment of the evidence, it may be easier to distort the results to one's liking than in the deterministic world of absolute representations of "the truth". For the risk analyst who encodes and processes the opinions of experts, honesty requires that he or she identifies and accounts for all reasonable hypotheses in the model so that the results involve all uncertainties and can be used accordingly. Therefore, the same problems of ethics, biases and distortion of knowledge to obtain questionable power, concern both the risk analyst and the expert.

Given these difficulties, Bayesian methods of risk analysis can therefore be considered "dangerous" ways of processing the available information. In the same way as wood cutting tools can be dangerous, the fact remains that these methods are irreplaceable as one of the few logical tools that allow treatment of uncertainties. They permit, in principle, ranking and comparing risks according to their magnitudes in order to determine which economic investments will allow protection of the largest number of people (Lave, 1987a), which in itself, is *a priori* a sound ethical goal. Yet again, there may be good moral and ethical reasons to adopt, perhaps at a cost to the safety of others, a different set of priorities, for instance to protect the weaker before the stronger or to eliminate dreaded threats before common ones.

9. Conclusions

Public perceptions of risks or their direct assessments by experts with strong positions at stake are not good guides to public decisions, nor are unfounded fears, or political maneuvers designed to permit ignoring the effects of one's decisions on others now or later. To begin to address the effects on risk management policies of human tendencies towards greed and ignorance, it is important to find out first what the evidence available today permits us to say about the magnitude of existing and potential risks.

One can then quantify the risk reduction benefits, for people today and in the future, that can be achieved by well managed expenditures given other investment possibilities.

To that end, investments in basic research are needed. But in order to address potential problems in a timely manner, incomplete information often has to be used and interpreted well before all scientific uncertainties are fully resolved. When the individual risk remains in the tolerable range, economics (costs and benefits) of risk mitigation matter because we are not infinitely rich and finite resources have to be spent wisely. Therefore, although it is not the only decision criterion, risk magnitude matters. The development of quantitative, probabilistic methods permits addressing risk management problems and setting priorities. These engineering methods permit identification and reinforcement of the weaknesses of a system (including the effects of human and organizational errors) and cost-effective allocation of risk mitigation resources. Obviously, the use of these mathematical techniques does not guarantee that all decisions will be enlightened and generous, but they provide a good place to start if those are the objectives.

Acknowledgment

The author thanks Dr. Vladimir Keilis-Borok and Dr. John Ahearn for their comments on earlier drafts of this paper.

REFERENCES

Alfimov, M., Corell, R., Courtillot, V., Intrilligator, M., and Keilis-Borok, V. (1997): 'Basic Science for the Survival of Humanity in the Third World War', *Kommersant Daily*, Nov. 29, 1997, Moscow, Russia.

Apostolakis, G. (1990): 'The Concept of Probability in Safety Assessments of Technological Systems', *Science*, vol. 250, December 1990, pp. 1359-1364.

Arrow, K.J. (1963): *Social Choices and Individual Values* (Wiley, New York, 2nd. Edition).

Budnitz, R.J., Apostolakis, G., Boore, D.M., Cluff, L.S., Coppersmith, K.J., Cornell, C.A., and Morris, P.A. (1998): 'Use of Technical Expert Panels: Applications to Probabilistic Seismic Hazard Analysis', *Risk Analysis*, vol. L8, no. 4, pp. 463-469.

Cornell, C.A. (1968): 'Engineering Seismic Risk Analysis', *Bulletin of the Seismological Society of America*, vol. 58, no. 5, October, 1968, pp. 1583-1606.

Cornell, C.A., and Newmark, N.M. (1978): 'On the Seismic Reliability of Nuclear Power Plants', Invited Paper, *Proceedings of ANS Topical Meeting on Probabilistic Reactor Safety*, Newport Beach, California, May 8-10, 1978.

Dalkey, N.C. (1967): *Delphi* (The RAND Corporation, P-30704, Santa Monica CA).

Davoudian, K., Wu, J., and Apostolakis, G. (1994): 'Incorporating Organizational Factors into Risk Assessment through the Analysis of Work Processes', *Reliability Engineering and System Safety*, vol. 45, pp. 85-105.

Ellsberg, D. (1961): 'Risk, Ambiguity, and the Savage Axioms', *The Quarterly Journal of Economics*, 75 (4), pp. 643-669, November.

de Finetti, B. (1974): *Theory of probability* (Wiley, New York).

Fischhoff, B., Lichtenstein, S., Slovic, P., Derby, S.L., and Keeney, R.L. (1981): *Acceptable Risk* (Cambridge University Press, New York).

Garrick, G.J. (1984): 'Recent Case Studies and Advancements in Probabilistic Risk Assessment', *Risk Analysis*, vol. 4, no. 4, pp. 267-279, December 1984.

Gell-Mann, M. (1994): *The Quark and the Jaguar* (Freeman & Co., New York).

Gleick, J. (1988): *Chaos: The Making of a New Science* (William Heinemann, London).

Howard, R.A. (1984): 'Value of Information Lotteries', in *Readings in the Principles and Practice of Decision Analysis* (vol. 2, p. 785), Strategic Decisions Group, Menlo Park California.

Henrion, M., and Fischhoff, B. (1986): 'Assessing Uncertainties in Physical Constants'. *American Journal of Physics*, vol. 54, pp. 791-798.

Intergovernmental Panel on Climate Change (1996): *Climate Change 1995. The science of Climate Change. The Second Assessment Report of the IPCC*, volume 1, J.J. Houghton *et al.* (eds.) (Cambridge University Press, Cambridge, U.K.).

Kaplan, S., and Garrick, B.J. (1981): 'On the Quantitative Definition of Risk', *Risk Analysis*, vol. 1, no. 1, pp. 11-27.

Keilis-Borok, V.I. (1972): *Computational Seismology* (Consultants Bureau, New York).

Keilis-Borok, V.I. (1980): 'Bursts of Seismicity as Long-term Precursors of Strong Earthquakes', *Journal of Geophysical Research*, vol. 85, pp. 803-811, Feb. 10, 1980.

Lave, L.B. (1987a): *Risk Assessment and Management* (Plenum Press, New York).

Lave, L.B. (1987b): 'Health and Safety Risk Analysis: Information for Better Decisions', *Science*, vol. 236, pp. 291-295, April 17, 1987.

Morgan, M.G. (1992): 'Prudent Avoidance', *Public Utility Forthnightly*, March 15, 1992.

Morgan, M.G. *et al.* (1995): 'Subjective Judgment by Climate Experts', *Environmental Science and Technology*, vol. 29, no. 10, pp. 468-476.

Morris, P.A. (1977): 'Combining Expert Judgment: A Bayesian Approach', *Management Science*, vol. 23, no. 7.

Nicolis, G., and Prigogine, I. (1989): *Exploring Complexity: an Introduction* (Freeman & Co., New York).

Paté, M.E., and Shah, H.C. (1979): 'Public Policy Issues: Earthquake Prediction', *Bulletin of the Seismological Society of America*, vol. 69, no. 5, October 1979, pp. 1533-1547.

Paté-Cornell, M.E. (1983a): 'Acceptable Decision Processes and Acceptable Risks in Public Sector Regulations', *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-13, no. 3, pp. 113-124, March-April 1983a.

Paté-Cornell, M.E. (1983b): 'Discounting in Risk Analysis: Capital vs. Human Safety', in M. Grigoriu (ed.), *Risk, Structural Engineering and Human Error* (University of Waterloo, Canada, 1983).

Paté-Cornell, M.E., and Neu, J. (1985): 'Warning Systems and Defense Policy: A Reliability Model for the Command and Control of the U.S. Nuclear Forces', *Risk Analysis*, vol. 5, no. 2, June 1985, pp. 121-138.

Paté-Cornell, M.E. (1986a): 'Warning Systems in Risk Management', *Risk Analysis*, vol. 5, no. 2, June 1986, pp. 223-234.

Paté-Cornell, M.E. (1986b): 'Risk Cost for New Dams: Economic Analysis and Effects of Monitoring', *Water Resources Research*, vol. 22, no. 1, January 1986, pp. 5-14.

Paté-Cornell, M.E. (1990): 'Organizational Aspects of Engineering System Safety: The Case of Offshore Platforms', *Science*, vol. 250, November 1990, pp. 1210-1217.

Paté-Cornell, M.E. (1994): 'Quantitative Safety Goals for Risk Management of Industrial Facilities', *Structural Safety*, vol. 13, no. 3, pp. 145-157, 1994.

Paté-Cornell, M.E., and Fischbeck, P.S. (1994): 'Risk Management for the Tiles of the Space Shuttle', *Interfaces*, vol. 24, pp. 64-86, January-February 1994.

Paté-Cornell, M.E., and Fischbeck, P.S. (1995): 'Probabilistic Interpretation of Command and Control Signals: Bayesian Updating of the Probability of Nuclear Attack', *Reliability Engineering and System Safety*, vol. 47, no. 1, pp. 27-36.

Paté-Cornell, M.E. (1996a): 'Uncertainties in Global Climate Change Estimates', *Climatic Change*, vol. 33, pp. 145-149.

Paté-Cornell, M.E. (1996b): 'Uncertainties in Risk Analysis: Six Levels of Treatment', *Reliability Engineering and System Safety*, vol. 54, pp. 95-111, 1996.

Paté-Cornell, M.E., Murphy, D.M., Lakats, L.M., and Gaba, D.M. (1996): 'Patient Risk in Anesthesia: Probabilistic Risk Analysis, Management Effects and Improvements', *Annals of Operations Research*, vol. 67, pp. 211-233.

Paté-Cornell, M.E. (1999): 'Conditional Uncertainty Analysis and Implications For Decision Making: The Case of the Waste Isolation Pilot Plant', *Risk Analysis* (in press).

Perrow, C. (1984): *Normal Accidents* (Basic Books, New York).

Popper, K.R. (1962): *Conjectures and Refutations* (Basic Books, New York).

Samuelson, P.A. (1976): *Economics* (10th Edition, McGraw-Hill, New York).

Savage, L.J. (1954): *The Foundations of Statistics* (Weiley, New York).

Serres, M. (1997): 'Le Serment du Scientifique', *Le Trésor* (Flammarion, Paris).

Slovic, P., Fischhoff, B., and Lichtenstein, S. (1982): 'Facts vs. Fears: Understanding Perceived Risks', *Judgment Under Uncertainty*, Khaneman, Slovic and Tversky (eds.) (Cambridge University Press, Cambridge, England, 1982).

Slovic, P. (1987): 'Perception of Risk', *Science*, vol. 236, pp. 280-285.

Starr, C. (1969): 'Social Benefit vs. Technological Risk', *Science*, vol. 165, pp. 1232-1239.

Tverski, A., and Khaneman, D. (1974): 'Judgment Under Uncertainty: Heuristics and Biases', *Science*, vol. 185, pp. 1124-1131.

USNRC (1975): Rasmussen, Norman C., *et al.*, *Reactor Safety Study: An Assessment of Accidental Risks in U.S. Commercial Nuclear Power Plants*, Nuclear Regulatory Commission, NUREG-75/014 (Wash-1400).

Winkler, R.L. (1986): 'Expert Resolution', *Management Science*, vol. 32, no. 3.

TOWARDS ASSESSING THE STABILITY AND SUSTAINABILITY OF COMPLEX SOCIO-ECONOMIC URBAN SYSTEMS

(A BLUEPRINT, A DAYDREAM, OR A NIGHTMARE?)

GASTON SCHABER*

1.

Human population as a whole is still in rapid growth (though at different speeds in different places) and is in the process of massively concentrating in cities and urban areas. Using crude estimates for the proportion of *people living in urban settings* at different points in time, we may say that 150 years ago they were one tenth; now, in global terms, they represent one half (but in the United States of America three quarters); in fifty years from now they may globally represent about nine tenths. *The cities as such* may mostly stand for social development, economic growth, and correspondingly for growing opportunities, but they may stand as well for poverty, deterioration of living conditions, social and economic exclusion, and growing disparities, growing levels of tension, and growing risks ...

Cities have been studied by social science scholars for decades, mostly at an individual level, *but it is only recently that cities have become a focal concern for national governments and for major international institutions and organizations.* These institutions now insist, when articulating urban concerns, that *much information on cities is already available but that it is not very enlightening, and is fragmented and unrelated*, reflecting the barriers (and interests!) which more often than not separate the actors and analysts involved, whether engineers and other professionals, or scholars, or business people, or politicians or just ... people.

---

* For the speaker's working perspective and base see note [1] at the end.

2.
Looking at Data and Indicators

Efforts are being made to overcome fragmented information and poor understanding by

— making critical use of existing data and collecting new information on urban conditions, change and trends;
— developing new analytical and modelling techniques;
— creating a knowledge base that should be helpful in devising and assessing urban programs and policies,

but these efforts are slow to bear fruit,
as shown by the following references.

Reference 1: United Nations Centre for Human Settlements

The United Nations [Contact at UNCHS (Habitat)] is running a *Global Urban Observatory*, Monitoring Human Settlements with Urban Indicators. The Draft Guide 1997 offers:

A Basic List of Key Indicators:

nine indicators for background data: on land use; city population; population growth rate; woman-headed households; average household size; household formation rate; income distribution; city product per person; tenure type, and

forty indicators for 6 modules:

1 socio-economic development (9);
2 infrastructure (4);
3 transport (4);
4 environmental management (5);
5 local government (8);
6 housing (10).

Module 1 comprises indicators behind the calculation of the *Human Development Index* (of UNDP) *at the City Level* – which in turn can be compared to the Human Development Index of the respective country.

* And a List of Extensive Indicators
which may be used to provide a fuller description.

More than 200 cities from over 100 countries have provided data on key indicators for a global database set up by UNCHS, Habitat.

The Data Are Now Available on: http://www.urbanobservatory.org

Reference 2: European Commission Documents

The European Commission decided on 15 June 1994 to establish a Community initiative on urban areas, with the acronym URBAN.

2.1. *The Urban Audit*

The call for tenders issued in this perspective by Directorate General XVI in June 1997 asked for

— the development of a methodology to deliver a maximum of data relating to a given set of urban indicators;
— the testing of the methodology on 58 European cities (on a given list);
— the coming out for each of the cities of an urban audit in a form that should be useable by the city authorities for subsequent updates;
— the examination of the 58 cities on each of the indicators developed in the project.

The Urban Audit should be a useful tool for urban policy-making:

— for giving an account on the state of affairs of the city, with a corresponding diagnosis;
— *for producing specific information on possible "pockets of risk" in relation to important disparities*;
— for giving access to comparative information on the other European cities in the project;
— for establishing exchanges with cities *sharing the same type of difficulties, with a view to finding joint solutions.*

The ultimate goal of the project is to produce a tool which the cities can use for self-analysis and self-evaluation.

Work may be done *down to the level of city districts and neighbourhoods.* At this small scale work does not have to be exhaustive, however it should identify *disparities* important enough to matter in terms of the socio-economic cohesion of the city. Distortions should be assessed with reference to mean values and observation should "zoom in" on problem districts. Cities should be compared from the point of view of their internal equilibrium.

Indicators: There is a Total of 34, Covering 5 Domains

Module 1 socio-economic aspects (18);
2 citizenship (3);
3 level of education and training (4);
4 environment (6);
5 leisure and culture (3).

Module 1 comprises some items contained in the UN module 6 (housing)

Comments:

— no reference is made to the UN initiative on urban Indicators, but;
— direct reference is made to possible pockets of risk related to important disparities (an aspect less explicitly addressed in the UN initiative);
— explicit reference to difficulties which may be similar for many cities and call for common problem solving.

THE EXPERIENCE IS STILL GOING ON/DATA NOT YET AVAILABLE
at the date of this presentation.

As soon as they are available data will be found at
http://www.inforegio.cec.eu.int/urban/audit

2.2. "*The City of Tomorrow and the Cultural Heritage*"
Key Action in the 5th Framework Programme of the EU.

Mention of this action is made here for the record, for two reasons: (a) the action is being launched only now, (b) in relation to our topic, the possible links are likely to be much looser than with the Urban Audit Project mentioned above.

REFERENCE 3: THE 1998 URBAN RESEARCH INITIATIVE (URI) OF THE US NATIONAL SCIENCE FOUNDATION

See the NSF homepage [http://www.nsf.gov/cgi-bin/getpub?sbe981]

The Initiative Announcement calls for *interdisciplinary research* on the *Dynamics of Change in Urban Environments*: "Research supported within URI is expected to facilitate the development of a predictive understanding of the complex interactions of people, the natural environment, and the physical settings of urban environments".

The underlying report "Understanding Urban Interactions" is authored by a group of 28 scientists, (with the Final Draft by Duncan M. Brown, mailto: dmbrown@crosslink.net). After stating that for multiple reasons "public and private decisions in urban matters often lead to solutions that are socially, physically, or environmentally less than optimal in the long run and on the larger scale", the group *focuses on one central question: Can urban systems be developed sustainably?*

and on four subquestions:

* How can the viability and sustainability of urban areas be defined, on scales ranging from the metropolitan to the community or neighborhood?

* What is required to produce strong and adaptable social capital and physical infrastructure in cities able to withstand change, hazards and disruption?
* What factors – technological, infrastructural, institutional, educational, social, behavioral, and informational – determine the political participation of urban citizens?
* What are the characteristics of life and institutions in extreme social environments?

To answer these questions, new tools and methods are called for:

— a truly multidisciplinary approach,
— using improved statistical and modelling techniques, and
— new systems for gathering and processing data,
— to address issues in greater complexity, and on a variety of scales.

Which tools and ways for modelling urban systems?

Tools and ways that:

— help overcome purely discipline-bound approaches which obscure rather than uncover relationships;
— make compatible the languages, data, models and findings of social scientists, engineers, and ecologists;
— look critically at the information available and identify intersections and boundaries where fruitful clusters of multidisciplinary research problems can be found;
— use a fresh systems approach, facilitated by new technology (database standards, inter-operating information systems, distributed computation, the Internet, spatial decision support systems, remotely sensed data, and GIS systems);
— cover wider ranges of geographical and temporal scales, from neighbourhood to metropolis and beyond, and into daily, seasonal and long-term rhythms.

Particular consideration should be given to:

* building detailed longitudinal data sets on small scales (census tract and neighbourhood cluster) on social and environmental indicators, and
* looking for parallel longitudinal data sets on the same scale as other social and ecological indicators related to income, education, and crime,

to make longitudinal correlations with census data on the same small scale.

Comments:

— no specific indicators are suggested in the "Initiative Announcement", in contrast to the United Nations' and the European Union's "initiatives";

— no reference is made to the UN or other non-US initiatives, except the observation that international comparisons are likely to be increasingly important and that international research agencies, private foundations, and firms should be interested in supporting comparative international studies on the subject; and no budget line is planned for participation in such an international comparative venture;
— the report insists on the importance of interdisciplinary, integrative approaches, and the importance of micro data interrelated over time and across space, on different scales, and down to the census tract and the neighbourhood;

as regards *extreme social environments*, the URI refers to them in a *low key manner* (this contrasts with the fact that even inner districts of such prestigious cities as Boston, Baltimore, or Philadelphia have "Human Development Index" values at Third World levels).

"Social scientists have many questions about the extreme social conditions of some urban communities (in which inadequate social capital and educational and political institutions and poverty are accompanied by a degraded physical environment). The determinants of social stability are poorly characterized at the neighbourhood level".

— "How can destructive environments be defined and identified through the interactions of social capital, physical infrastructure, and political, economic, and ecological conditions?".
— "Is the analogy of disaster useful? (Can social early warning systems be imagined?)".

DATA ARE NOT YET AVAILABLE: THE PROGRAMME IS JUST STARTING.

REFERENCE 4: THE APPROACH IN MONITORING AND EVALUATING URBAN POLICY OF THE FRENCH NATIONAL INSTITUTE FOR STATISTICS AND ECONOMIC STUDIES (INSEE = Institut National de la Statistique et des Etudes Economiques).

Document by Pascal OGER, 1995, INSEE-CENTRE, ORLEANS: "INSEE's involvement in the monitoring and evaluation of urban Policy in France: KEY SOCIO-DEMOGRAPHIC INDICATORS FOR CITY DISTRICTS".

Here are the main characteristics of this one-country project, *which focuses very clearly on the issues of our subject*: it provides data to be used for monitoring and evaluating urban policy, related to *problem areas* to be *seen in a larger context*; it is tested in the 8 conurbations of the Central

Region, for each urban district ("*quartier*") it makes *annual calculations* for *socio-economic indicators* related to the major aspects of the urban policy defined. The indicators are calculated in a strictly identical manner to guarantee comparability in space and over time for all the districts, and they are updated annually in the same way for the duration of the Plan Contract.

The sources of indicators: they come mainly from administrative files. e.g. from the Family Allowances Offices, the National Employment Agency, statistical services of various ministries, and from local questioning of institutional officials, such as school heads or managers of various agencies, and of course from INSEE's in-house production (e.g. on economic activity).

KEY INDICATORS

Module A: Population-settlement patterns (25);
B: Poverty and risk of poverty (9);
C: Residential environment and housing (7);
D: Education and training, pre-primary (3), primary (6), secondary (11);
E: Employment and unemployment (14);
F: Economic activity (enterprises) (4);
G: Civic life;
H: Delinquency.

Comments:

— a very interesting experimental initiative, drawing on all the resources, and on the legal statute, of INSEE, which is authorized to make use of individual files of administrative origin;
— definitely aware of the methodological difficulties at all the levels involved;
— working to extend at the district level the existing modules and to add new ones (health, culture, environment ...);
— working on the intricacies of including at the level of districts data on delinquency from the national police records.

Here INSEE deals, within the context of very well documented and highly relevant domains (A to G), with an important issue: assessing and scaling critical neighbourhoods for (H): vandalism, group delinquency, aggressive behaviour (from verbal attacks to physical violence), as well as for spells of collective clashes and for riots. (Refer in this context to the work done at the Ministry of the Interior by Lucienne BUI-TRONG [2]). Information and insights will be produced from the micro-data, processed and published under the required French rules for confidentiality, which will certainly be relevant to our concerns.

But Will There be Sets of Anonymized Microdata Available for the Scientific Community?

At this point we may summarily reflect upon our four references and the wearisome inventory of domains and indicators I offered to your attention. The list of items, whether from the world-wide project of the Unites Nations (initially shared by the World Bank), or the almost continent-wide project by the European Union, or the national projects of the United States and of France, are very similar, though they reflect some interesting differences, even at the level of intentions and concerns. But the most striking aspect they have in common is their ... lack of relatedness! They are not conceived in a perspective of reciprocity or complementarity! In the case of one of the countries involved, I had the opportunity to introduce two research teams working for years on strongly related issues without being aware of each other's which had efforts and results ...

3.

While Waiting for City Data and Indicators Still to Come out of These Huge Initiatives,

and while a waiting that by some wise but unlikely decision these data from important international or national organizations might be coordinated,

what could/should *we* do?

By "we" I refer to a small number of research institutions and their networks who did not wait for the larger UN or US or EU initiatives to produce and process relevant micro-data for comparative research purposes, and who are giving to the scientific community working in the socio-economic domains protected access to their anonymized data sets. (For access to information on such research institutions and networks, see our own Center CEPS/INSTEAD's homepage: http://www.ceps.lu). There are scarce, but valuable opportunities already offered to researchers at the data level; for our Center's comparative data bases see [3].

3.1. We should deliberately commit the research potential and resources *we ourselves can decide about*,

* to work with partner institutions and networks prepared and willing to cooperate in a truly interdisciplinary way, and in a comparative perspective;
* on micro-data sets, preferably longitudinal ones, for dealing with basic issues of socio-economic change over time and across space;

* taking care that our joint approaches stay integrative, i.e. able to deal with complex settings, factors, and interactions;
* to make common projects, with defined objectives and tasks, means and strategies, with high standards for data caring and data sharing;
* in order to make a stimulating contribution to the development of a badly needed research infrastructure in the socio-economic sciences.

3.2. As regards the *data situation* and *accessibility* for individual researchers:

3.2.1. we already have available amounts of longitudinal socio-economic data on income and *living conditions of persons and households*, data which to a large extent are already comparable, and even thoroughly documented (see our Center's homepage for information on such databases). Many of these data sets are from one-time surveys but they have already been put together in time series; and they are complemented by a growing number of longitudinal studies which have *information about situations, events, and changes in the lives of thousands of people* (each individual being specified by numerous variables year by year and in some of these panel studies for more than 10 or even 20 years) and which have enormous analytical pover in comparison to other data sources.

Both for persons and for households or families, *these events and changes*: may mean "transitions" and "transformations" of many kinds, may be slow, fast, steady, irregular, abrupt, dramatic may be positive, negative, critical, catastrophic, and may mean "spells" of poverty, dependency, unemployment, sickness, etc.

In the context of the present study-week, I would like to mention some challenging issues that could be dealt with *by using these databases*:

— with regard to *methodology*: all these data about changes could be approached in alternative ways, *linear and non-linear* ...
— with regard to *basic questions*, e.g., about the demographic and social texture:
there have never been so many generations alive at the same time,
there have never been so many middle-aged couples having more living parents
than children,
there have never been so many children having so many parents to refer to (or not refer to),

all this being given, what are the relationships between all these people and all these generations? What are the exchanges (and direction of exchanges) in the major dimensions to be measured: time, money, shelter? What is the

proportion of the total population for which the possibility is given or not given to have such intergenerational interaction? These are just some examples of topics akin to our concerns. It would be easy to extend the list (child poverty, longevity and work life ...). All this is in regard to our *social* tissue. What about the *economic* tissue and fabric?

3.2.2. As for the potential of firms and businesses data, we could make similar comments and considerations, but not without two important restrictions: longitudinal studies on firms are much less developed, this is particularly true of comparative ones. Where they exist they are mostly run by official statistical offices and hardly accessible to the scientific community (except for special arrangements, such as those taken by the US Bureau of the Census and its Safe Data Center Arrangements with selected Universities).

3.2.3. Despite these restrictions and obstacles, some efforts are already being made *to bring together information derived from longitudinal studies on people and from similar studies on firms* – in settings which are controlled for data protection and confidentiality, mainly in the US and in the Scandinavian countries – producing valuable insights regarding both social and the economic fabrics and dynamics.

3.2.4. At this point, I would like to make a brief remark, which has importance both for epistemological and for practical reasons, precisely in relation to the specific topic of cities and complex urban settings we are dealing with now. Until now we have made use of *two basic units of observation*: (a) the *person* for studying the social tissue, (b) the *firm* for studying the economic tissue – both being followed over time. *But we need a third one*, (c) an *area unit*, to study them also in space.

We will opt for the smallest unit which is most instrumental for studying complex socio-economic systems: *the neighbourhood or district* (in French the "*quartier*") – in spite of all the difficulties one encounters when dealing for comparative purposes with a unit that by conception presents variable shapes and dimensions and, moreover, is not necessarily an administrative unit.

Technically we need the neighbouhood as a convenient unit for micro-mapping, but we need it even more because it is a living space where multiple interrelated characteristics, impacts and actions shape an environment that matters for its residents in many respects.

Most of the longitudinal studies that we refer to under 3.2.1. have established so as to be representative of a given country's population (or a subset, like youth), *but not necessarily to be representative in a geographical sense*. And the studies at the beginning did not "attach" to the individuals

in the sample some characteristics of the environments they come from or stay in. Partly for reasons of confidentiality as well, the scientific use files accessible to researchers often contain only the distinction urban/rural, without other geographical specification.

Now, some of these longitudinal studies, particularly in relation to the children involved, *proceed to some "geo-referencing" and bring in particular neighbourhood variables obtained from other sources and "pull them over" the corpus of information already gathered* – which is a way of examining the possible influence of the near environment [4]. However tricky this may be, it is an interesting approach by which to bring potential neighbourhood effects into focus. And we should continue to use it since it will still take considerable time and effort before we will be able to look at these complex interactions *"the other way around", i.e. starting with a thorough study of the different kinds of neighbourhoods*: a basic part of what the new urban studies should probably be about.

4.
Neighbourhoods in Their Urban Context
Focal Areas of Concern/Focal Areas of Risk

In 3.2.4. we were looking for methodological arguments to choose the neighbourhood as the basic area unit for studying complex socio-economic systems. These arguments are also supported for policy reasons: we will have to look scientifically into neighbourhoods because they are becoming focal areas of national and international policy concerns, on which independent consultancy and assessment work will be required.

The neighbourhood concept surges up in a very straightforward way in a major European Commission document, which I would like to quote here. I know that it is unusual, even in an informal academic exercise, to incorporate an official text into a presentation. But the text is so concise and to the point that it would be a mistake just to paraphrase it. So this notice of the Commission to the Member States will be reproduced in annex [5]. Here a short excerpt: "Some of the Community's most acute problems associated with lack of economic opportunity, low incomes and a generally poor quality of life are found in urban areas. The growing tensions within European society are evident particularly in the serious level of social exclusion in an increasing number of inner city or peripheral urban areas".

5.

Summing up

What did we learn while going through these major UN-, US-, EC- initiatives?

* Contemporary society is massively concentrating in cities and urban areas, where it generates the conditions and the push for equally massive wealth, progress and change.
* But the urban world it creates, stands not only for social development, economic growth and corresponding opportunities,
— it also stands for growing inequality, poverty, deterioration of living conditions, and social and economic exclusion,
— and growing disparities, tensions, and risks (of unrest, violence, disruption ...).

So much for the basic concerns to be dealt with in cities, urban areas and neighbourhoods ...

* Insofar as these initiatives are addressed to potential applicants, they give information on how to proceed. And also on key indicators (UN, EU), – except for, the US-URI, which on the other side insists strongly (1) on new approaches which are integrated, interdisciplinary, and (2) on the necessity to collect and use longitudinal microdata, (3) whenever possible geo-referenced, and (4) at various scales.

6.

Where do we Stay, Where do we Go? "We" Being our Centre and our Partners

6.1. For two decades our energy and resources all went into
— the development of *socio-economic micro-databases built on nationally representative surveys and longitudinal panel studies*;
— which have to be continuously documented, both technically and institutionally;
— to be made comparable;
— to be updated, and managed, and kept available for the research community.

[*Nota bene*: databases can reach higher quality only when a sufficient number of high quality users of different scientific orientation help you through their critical and constructive remarks to make them better ...].

6.2. For more than a decade we have also been working (very, very slowly for lack of means) at developing – as always in comparative and longitudinal perspectives – *an integrated and geo-referenced information system for multi-dimensional data*, at the level of *cities* and their *regions*, in the context of massive regional *structural change*.

Funding is here not the only problem. For any research of this type, and at these meso- and micro-scales, it is equally crucial and difficult *to get access to relevant microdata produced by local and/or regional, national administrations* – for these research purposes publicly available data do not suffice. A research centre has to prove that its independence (scientific, legal and other), and its standards for data confidentiality and data protection are such that it will be authorized to access and process sensitive administrative data [*trust building capability*]. And the centre must be prepared *to offer services to these administrations in return for data access*: e.g. to help them to set up internally a data management system which they can take over for their own administrative purposes [*negotiation capability*]. These processes take time.

6.3. In October 1993, in Atlanta, Georgia, at a CIESIN [6] Workshop aimed at *integrating human/social data into CIESIN's Earth Science Information Network*, the present speaker submitted to the organizers a proposal for a *pilot-project to develop an integrated information system*

* designed to *combine, in a meaningful way, demographic, social, economic and ecological data of mutual relevance*, for the purpose both of comparative research and policy analysis;
* bringing together in a *multi-dimensional mapping system*
  — selected demographic, social, economic and environmental variables,
  — presented at different levels of (dis)aggregation, including for each country or region the lowest registration level obtainable, and
  — in the context of the respective policy issues, choices and regulations (local/global; domestic/crossnational);
* *documenting the information status of these heterogeneous sets of data and measures*, and *assessing their respective value for combined analysis* of the complex interrelationships between physical, socio-economic, technological, cultural and political factors, issues and policies.

No funds were available in the CIESIN budget for a comparative transatlantic pilot-project of this kind.

6.4. In the same year (1993), Keilis-Borok and Schaber met in INTAS, a European Commission supported International Association for Promot-

ing Cooperation with Scientists from the Independent States of the Former Soviet Union. There Keilis-Borok and Schaber had the opportunity to find out that professionally they were both dealing with "RISK", the one with "EARTH", the other with "PEOPLE" and that some similarities in their approaches were intriguing enough to make them both think about new forms of transdisciplinary cooperation ...

By September 1994 financial support from the Luxembourg Ministry of Foreign Affairs for a bilateral project of cooperation with Russian scientists helped the new partners MITPAN (Keilis-Borok) and CEPS/INSTEAD (Schaber) to prepare a pilot-project centered on the development of indicators for assessing stability/instability of complex social systems in a comparative perspective ... It is still under development and is aimed at monitoring and possibly anticipating socio-economic developments (steady trends and crises) by the analysis of relevant and available data bases – combining the data base and socio-economic expertise of CEPS/INSTEAD with the mathematical expertise of MITPAN in the prediction of highly complex chaotic systems, political and economic phenomena included.

6.5. There is one more module to be integrated: knowing about our centre's activities in building and managing comparative databases and in making these accessible to the research community, a network of researchers coordinated by Prof. Wolfgang Voges [7], of the University of Bremen, has recently asked CEPS/INSTEAD to host *their comparative database on welfare regimes and welfare use in 12 selected cities in 6 European countries (D,E,F,I,P,S) and to include it for users in our Large-Scale Facility programme.* This longitudinal database is a candidate for being a major sub-module for assessing important aspects of socio-economic performance of cities down to the neighbourhood level:

* It carries e.g. anonymized democraphic information on individuals defined as poor and entitled to benefits according to the respective local rules;
* longitudinal information about individuals' recipiency of benefit (level and duration), and reasons for entitlement;
* information on the social and economic characteristics of the neighbourhood the recipients live in, related to corresponding data for all neighbourhoods of the city.

The potential for analysis is considerable, as for modelling and simulation purposes: what would be the situation of people in need if the authorities of their City X would apply the regulations and criteria of City Y? And vice versa?

We plan of course to find means and ways by which to enlarge the

scope of this database and motivate additional cities to join the project (and why not a US one?). – And since some of the cities included in the present database are also in the UN Global Urban Observatory and/or in the European Urban Audit, there should be opportunities for a few complementary comparisons ...

While closing this section on "where do we stay and where do we go?", we see no reason to reach out for an overall conclusion. We can only state that we are committed to the task of building elements (methodology-related and issue-related as those we presented above) which should contribute to developing, in the socio-economic sciences, a research infrastructure able to deal with the concerns we have just outlined ...

The framework for this work to be done:

* our common research networks: http://www.ceps.lu
* our Integrated Research Infrastructure in the Socio-Economic Sciences at CEPS/INSTEAD, IRISS-C/I: http://www.ceps.lu/iriss/iriss.htm

NOTES and REFERENCES

[1] About the speaker's work (a) and work base (b):

a.1. clinical psychologist working in the field of people (young and adult) who runthe risk of poverty, deviancy, or criminality, and with people representing themselves high risks (offenders sentenced for homicide, rape, etc.),
a.2. and progressively combining clinical and quantitative approaches in these fields,
a.3. and being led to put his psychological and his social work into a broader context by the description and analysis of socio-economic conditions and corresponding public policies,
a.4. which in turn led him (since 1978, and initially in the context of the First European Program to Combat Poverty) to create interdisciplinary teams of researchers to develop large sets of socio-economic micro-data built up from one-time surveys and from longitudinal panels, in an internationally comparative perspective,
a.5. in order to work out a broader Monitoring System for Social and Economic Policy (description, analysis, simulation/modeling, programming ...).

b) work base:

At our Centre we develop large sets of socio-economic micro data: national and international comparative databases for research and for public policy monitoring.
Our micro-databases relate to
b.1. Income Distribution, Labor Force Participation, the Living Conditions of People (individuals/households) - as well as

b.2. FIRMS, BUSINESSES and SERVICES (although the latter data sets are still considerably less well developed than the former) - as well as
b.3. data serving to develop INTEGRATED INFORMATION AND INTELLIGENCE SYSTEMS on demographic, economic, social, and ecological issues and change (at a local level first, before moving to regional, national, international levels). This programme is still seriously underdeveloped.
* The micro data come from a variety of sources: one-time national surveys (put into time series whenever possible); national longitudinal studies which follow the same units of observation over years; official statistical or administrative records when legally available.
Providers of the data are, according to country arrangements, national science foundations, or other academic institutions, or national statistical offices.
* Our databases do not just contain variables and figures: they are documented with care for the researchers,
— at the technical level for unambiguous use, and
— at the institutional level (national laws, regulations etc.) for a better understanding of what the figures and numbers may mean and stand for.
* Production and management of the databases as well as access offered to the scientific user community take place under legally defined conditions and rules for data protection and confidentiality.

[2] BUI-TRONG, Lucienne, 1993, "L'insécurité des quartiers sensibles: une échelle d'évaluation", Les Cahiers de la Sécurité Intérieure, n° 14, Paris.
— 1998, "L'échelle de la violence urbaine; de l'aide à la decision a la gestion des risques", Les Cahiers de la Sécurité Intérieure, n° 33, Paris.

[3] The CEPS/INSTEAD comparative databases LIS, LES, PACO:

* The "Luxembourg Income Study" (LIS), since 1983.

LIS is an international comparative cross-sectional database on income distribution which presently contains 80 micro-data sets from 26 countries covering the period 1968 to 1997: Australia, Austria, Belgium, Canada, Czech Republic, Denmark, Finland, France, Germany, Hungary, Ireland, Israel, Italy, Luxembourg, Netherlands, Norway, Poland, Portugal, Russia, Slovakia, Spain, Sweden, Switzerland, Taiwan, United Kingdom, United States.

The database is accessed globally via electronic mail networks by over 300 users in 28 countries (the jobs being sent are closely checked for reasons of data protection). In addition to harmonized data, the LIS users are offered extensive documentation at the technical as well on the institutional level.

* The "Luxembourg Employment Study" (LES), since 1993.

The LES is an international comparative database integrating microdata from a set of Labour Force Surveys from the 1990's. The database presently includes datasets from 15 countries, with about 90 variables, new datasets being continuously added. The countries included are: Austria, Canada, Czech Republic, Finland, France, Hungary, Luxembourg, Norway, Poland, Slovakia, Slovenia, Spain, Sweden, Switzerland, United Kingdom, United States. The following countries are to be entered in the near future: Australia, Belgium, Bulgaria, Denmark, Israel, Italy, Lithuania, Netherlands, Portugal, Rumania.

* The "Panel Comparability Project" (PACO), since 1990.

The PACO is an international Data Archive and Database integrating national longitudinal household panels from West and East. The PACO Data Archive presently includes 70 original panel datasets from 17 countries, with original (not standardised) variables transformed into a common format (SPSS system files for Windows on PC); the PACO Database contains scientific

use files and presently includes 40 datasets with fully comparable standardized variables from household panels of the following countries: France (Lorraine), Germany, Hungary, Luxembourg, Poland, United Kingdom, United States. The PACO network is still developing at CEPS/INSTEAD and will include panels from additional countries (Belgium, Russia, Sweden ...). Information in the PACO files is available (a) for households and individuals on the micro-level, (b) for single years, (c) in longitudinal form – completed by extensive technical and institutional documentation. The scientific use files are available on CD-Rom, according to specified confidentiality rules and pledges.

[4] See "Child Development in the Context of Family and Community Resources: An Agenda for National Data Collections", by Jeanne Brooks-Gunn, Brett Brown, Greg J. Duncan and Krisitn Anderson Moore, in "Integrating Federal Statistics on Children: Report of a Workshop", 1995, National Academy Press: http://books.nap.edu/books/0309052491/html/27.htm

[5] Here the Commission text, as published in the Official Journal of the European Communities, No C 180/6,1.7.1994:

Notice to the Member States

laying down guidelines for operational programmes which Member States are invited to establish in the framework of a Community initiative concerning urban areas

(URBAN)
(94/C 180/02)

1. At its meeting on June 15 1994, the Commission of the European Communities decided to establish a Community initiative concerning urban areas (hereafter called URBAN) ...
2. In the context of URBAN, Community assistance in the form of loans and grants and technical assistance is made available for measures in areas which respect the guidelines laid down in this notice, ...

I. Scope and Objective

3. Some of the Community's most acute problems associated with lack of economic opportunity, low incomes and a generally poor quality of life are found in urban areas. The growing tensions within European society are evident particularly in the serious level of social exclusion in an increasing number of inner city or peripheral urban areas.
4. Problems are often aggravated by the financial difficulties of many urban local authorities who are unable to provide increasingly expensive support services for a less affluent population. This leads to the deterioration of the urban fabric, the impossibility of renovating or replacing obsolete infrastructures and the disappearance or severe reduction of economic activity in the worst affected areas.
5. Difficult neighbourhoods within cities may be identified geographically. Certain socio-economic indicators are significantly worse than the average of the city or urban agglomeration area. These would include unemployment levels, education attainment, the crime rate, standard of housing, the percentage of social-welfare benefit recipients, the socio-ethnic mix, environmental decay, deterioration public transport and poor local facilities etc. These deprived areas can also be within generally prosperous cities, or in cities which are the most prosperous parts of a less developed region.
6. Urban issues should be tackled in an integrated way, supporting business creation, improving infrastructures and the physical environment, providing customised training, actions for equality of opportunities and social amenities. The combined effort of ERDF and ESF will be needed, but is has to be complemented by other resources.
7. This initiative cannot hope to match in scale what is in effect a major problem of contempo-

rary society. It aims instead to act as a catalyst in a broad-based approach, by undertaking key schemes to help deprived urban areas achieve a lasting improvement in living standards for their inhabitants.

8. It will provide assistance to the responsible authorities in their efforts to provide the necessary amenities so as to attract economic activity and create confidence and security for the population living in the areas, integrating them into the economy and social mainstream.

[6] CIESIN is a NASA related entity: Consortium for International Earth Science Information Network, 1825 K Street, N.W., Suite 805, Washington, D.C., 20006 USA.

[7] Dr Wolfgang Voges, Universität Bremen, Zentrum für Sozialpolitik, Parkstrasse 39, D-28209 Bremen, Tel: 0049 421 218 4367, Fax: 0049 421 218 4052,
E-mail: wovo@athene.barkhof.uni-bremen.de

THE NATURE OF CRITICAL TRANSITIONS IN SOLID EARTH (THE PROBLEMS OF THEIR MODELLING, PREDICTION AND CONTROL AND THEIR POTENTIAL IMPLICATIONS FOR SOCIO-ECONOMIC CRISES)

VLADIMIR I. KEILIS-BOROK

SUMMARY

The Lithosphere as a non-linear system which generates geological disasters.

Geological disasters – earthquakes, volcanic eruptions, landslides, and floods – are generated in the outer shell of solid Earth, called the *lithosphere*, the domain of rocks. Its thickness ranges from a few *kms* in the ocean ridges to a few hundred *kms* in certain continental regions.

The lithosphere is set in motion by the impact of the large scale currents in the underlying mantle of the earth and by a variety of internal processes. Two major properties of the lithosphere turn it into a hierarchical chaotic system.

— *The hierarchical, probably fractal, structure.* The lithosphere is divided into a hierarchy of volumes ("blocks"), which move relatively to each other. Larger blocks are tectonic plates with a linear dimension of 10_3 - 104 *km*. They are consecutively divided into smaller blocks, down to about 1,025 grains of rocks with a characteristic dimension of *cms*. The blocks are separated by boundary layers of a similar hierarchical structure, with a more dense division.

— *Complex instability.* The relative movement of the blocks is controlled, roughly speaking, by the {stress - strength} field. The stress, and particularly the strength, are controlled in turn by a multitude of processes, each causing strong instability: among them are mechanical and chemical rock-fluid interactions, phase transformation of the minerals, buckling,

micro fracturing, etc. Each of these mechanisms may trigger a drop in the lithosphere's strength by a factor up to 105. Except for rather special circumstances, none of them predominates so that the others can be neglected, and fundamental equations may be realistically developed at an elementary level. Altogether these mechanisms turn the lithosphere into a chaotic system of the type indicated above.

Major features of this system.

The study of this system is so far at the pre-equations stage – the phenomenological search for the major regularities necessary to develop an adequate theoretical model. This is essentially a pattern recognition problem. Important regularities of such a kind have been found in the observed and modeled seismicity generated in the lithosphere. About a million earthquakes are registered every year, a few hundred of them are destructive and once in a decade or two a devastating catastrophe occurs. The regular macro- and meso-scale features of the dynamics of seismicity relevant to our topic are summed up in Appendix 1. They include the symptoms of critical transition determined by pattern recognition of infrequent events – a methodology developed by the I. Gelfand school.

The measures of success in the modeling of this system include:

— Low - parametric reproduction of its major features listed in Table 1. Some of them are reproducible by such a diversity of models that the existence of a much simpler toy model seems possible.
— Obviously - new features, to be confirmed by observations.

The performance of earthquake prediction algorithms.

The problem of prediction is posed as a consecutive reduction of the time-space domain where a strong earthquake has to be expected. Several prediction algorithms are based on the features in Appendix 1. The level of averaging as defined by these features corresponds to intermediate-term prediction with a characteristic duration of alarms of a few years. On the average these algorithms allow us to predict about 70% of strong earthquakes with alarms occupying 10% to 20% of the space-time covered by prediction. Such predictions allow us to prevent a considerable part of the damage, though not the whole damage. The algorithms are tested by advance prediction in numerous regions worldwide. High statistical significance at a confidence level of 95% to 99% is established. Among those

predicted have been all of the last 7 great earthquakes with a magnitude 8 or more. To complete the project, a distribution of predictions to 60 leading scientists and administrators was recently effected. The test is unprecedented in rigor and volume.

What did we learn from the premonitory phenomena?

Obviously, they are pivotal for a fundamental understanding of the dynamics of the lithosphere as well as for disaster preparedness. Major findings reviewed here are summed up in Appendix 2; they are obtained by the "holistic" approach. The nature of premonitory phenomena (the first entry in Appendix 2) is discussed in more detail in the next section.

The nature of the critical transitions considered.

— *Universality.* Many features of the dynamics of the lithosphere (Appendix 1) are reproduced but so far not all on the lattice models which are not earth-specific but can be equally attributed to many systems of interacting elements. This suggests that we have probably encountered in seismicity the scenarios of critical transitions common to a wide class of chaotic systems. The decisive system-specific element in such a scenario is the definition of "control parameters" i.e. the features which reflect the critical transition considered. After they have been established (pattern recognised), transition becomes partly similar for different systems.

— *Multiple fracturing.* This similarity is relatively well established for multiple fracturing in exceedingly different conditions, from a neutron star through the lithosphere, geotechnical and engineering constructions down to laboratory samples of solid materials. The energy released by observed fractures ranges from 1041 erg for the starquakes down to a fraction of an erg for laboratory samples.

— *Socio-economic systems.* Two problems were considered for such systems: the prediction of economic recessions and the outcome of American (presidential and mid-term senatorial) elections. Pattern recognition of macro- and meso-scale parameters reflecting critical transition was unexpectedly successful in each problem, both for understanding the system and for prediction in a practical sense.

— *The limits of universality.* In the background of the "universal" features of critical transitions the system-specific features may be established. In the earthquake prediction problem they are determined by the geometry

of the system of blocks and faults comprising the lithosphere. So, the triggering of strong earthquakes is controlled in the fractured zones ("nodes") formed around fault intersections.

Emerging possibilities.

Macro- and meso-scale premonitory phenomena were considered here in the intentionally coarse robust definitions. As a start this is inevitable, making the first approximation to prediction stable and reliable at the price of limiting its accuracy. The following possibilities now emerge by which to develop further approximations to prediction:

— Renormalization of premonitory phenomena on consecutively smaller scales, as long as the observations are sufficiently complete for that purpose.
— Recognition of the changes in the type of criticality over time ("prediction of predictability").
— Incorporation of system-specific features into prediction; for example they may determine parameters of "universal" models.

For the earthquakes, a 5-fold rise of prediction accuracy with transition to short-term prediction may by reasonably expected.

The merger of mathematical modeling and existing data bases opens up considerable but still untapped possibilities of such a king.

Appendix 1.

Hypothetically Universal Symptoms of Critical Transitions

in hierarchical non-linear dissipative systems with "intermediate number of degrees of freedom.

— These symptoms are established by the analysis of observed and modeled seismicity.
— They are tested (successfully) only for seismicity, advance earthquake prediction.
— Their formulation is slightly generalized ("seismicity" is replaced by "activity") to discuss their implications for other disasters.

General Features

Power-law size distribution, $dN(E) \sim E\text{-}BdlgE$, $B \approx 5/3$,
in a limited range of E, with the downward bend on both sides.
This distribution (emerging after a gross averaging in time and space) is fundamental for non-linear dynamics; it is encountered in a multitude of processes.

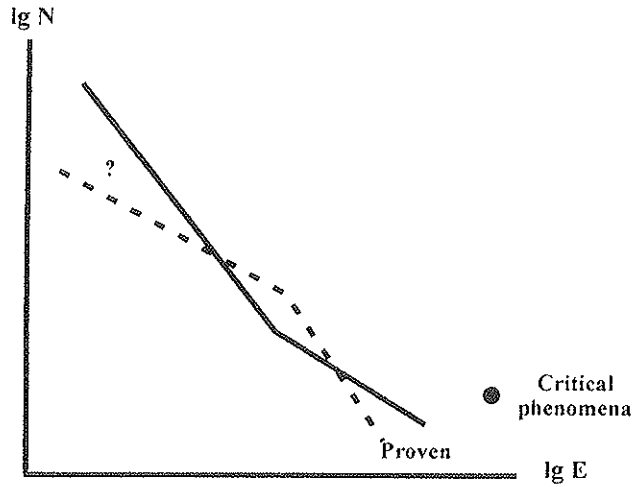
Hierarchical clustering in space and time.

These fundamental features are necessary but insufficient for the prediction and control of critical phenomena. They are reproduced in a broad variety of models. Non-universal is only the domination of aftershocks in the clusters.

Premonitory Transformation of the Background Activity of the System ("Static")

The following macro- and meso-scale characteristics of the static reflect the approach of the critical phenomenon:

— *Deviations of size distribution from the power law*



Solid line corresponds to approach of the break up.

— *Response to excitation (increasing)*
— *Size of the clusters; range of correlation in space; transient intensity; its irregularity in time. (all increases)*
— *Dimensionality (probably decreases)*

The tests under way: earthquake prediction, geoengineering instabilities, economic recessions, social instability in urban areas.

Appendix 2.

Two Complementary Approaches to Earthquake Prediction

— "reductionism" (from details to the whole)
— "holism" (from the whole to details)

Premonitory phenomena (PPH)

are specific to mechanisms, e.g. are divided into: friction, fluid filtration, stress corrosion, buckling, etc.
— "universal" ones common for many chaotic systems
— those, depending on geometry of block & faults system
— mechanism
— specific ones

PPH in different regions and energy ranges

— are different
— are to a considerable extent similar

PPH preceding an earthquake with linear source dimension L are formed

on a fault segment of linear size from L to 2L in a system of blocks and faults of the following linear scale:
— up to 50L in time scale tens of year
— from 3L to 10L in time scale years
— from L to 2L in time scale years or months

Constitutive equations

— are local
— are non-local

Nucleation of earthquakes is controlled

— by stress-strength relation on a fault segment
— also by the geometric incompatibility near the fault junctions which may supersede a stress-strenght criterion.

III.

SCENARIOS OF TRANSITIONS TO CRITICAL PHENOMENA

FRACTAL FINANCIAL FLUCTUATIONS; DO THEY THREATEN SUSTAINABILITY?

BENOIT B. MANDELBROT

> "Were all the fountains of the great deep broken up, and the windows of heaven were opened. And the rain was upon the earth forty days and forty nights" (*Genesis*: 6, 11-12).
>
> "There came seven years of great plenty throughout the land of Egypt. And there shall arise after them seven years of famine" (*Genesis*: 41, 29-30).

The very diversity of the speakers at this meeting suffices to show that sustainability is threatened by very diverse obstacles. Several speakers have described world-wide trends whose continuation would make sustainability impossible. My presentation is, instead, concerned with instances where the trends themselves might be tolerable but severe deviations from the norm are not, in fact, may well be unacceptable. For good reason, and as witnessed by the quotation from the Bible that opens this text and will be commented upon momentarily, a deep concern with those deviations has an extremely long history. Until recently, no scientific tools were available to begin to face those deviations, not even to measure their intensity. This paper will not discuss all the tools now available, only those the speaker has introduced very recently.

Based on the two quotations from the Bible, my terminology classifies the threatening deviations from norms in two categories. Because of the story of the Flood and Noah, I denote by the term "Noah Effect" those major changes that occur rapidly (even instantaneously) but have strong

and durable consequences. Because of the story of the dream of Pharaoh and Joseph the son of Jacob, I denote by the term "Joseph Effect" those sequences of changes that need not seem threatening when viewed individually but become major in their cumulative effect.

The Noah and the Joseph obstacles to sustainability range from the wholly natural (earthquakes, volcanoes, non man-made climate changes) to the wholly man-made. Both "effects" have long characterized the phenomena I have studied very energetically. Many of those studies were reported or summarized in my 1982 monograph, or included in my Selecta books, of which the first three have been published. These are big books and it would be of little use to try to summarize them in here. Instead, I shall focus on one topic and make minimal reference to the other aspects of the Noah and Joseph Effects.

That topic is the variation of financial prices, that is, of prices quoted on financial markets that trade securities, commodities, and exchange or interest rates. The wild volatility of those markets has long been known but only recently did it form an intimate link with such issues as the sustainable growth of developing or other economically fragile countries.

The details of that intimate link belong to political economy. They will be briefly touched upon at the end of this text, but to keep in focus and manage space, an extensive discussion will be carried out elsewhere. The sole ambition of this paper is to further the knowledge of the underlying facts. Unfortunately and perhaps surprisingly, the existing models of financial price variability for good enough and new research is needed. Such statements tend to be viewed as self-serving – and an excuse for delaying urgent action. But in the present instance skepticism would be unwarranted. New research is keenly needed, because the points that matter most in the discussion of smooth sustainability coincide very precisely with points that previous research had deliberately set aside or disregarded. The text that follows is a draft of Chapter 1 of a forthcoming book of mine.

This book touches many topics, but its main ambition is to contribute to a better understanding of price variation. Inevitably, it criticizes previously held views on this topic, particularly, the "coin tossing model", to be described momentarily.

The point of departure is that financial prices, including those of securities, commodities, foreign exchange or interest rates, are largely unpredictable. The best one can do is to evaluate the odds for or against some desired or feared outcomes, the most extreme being "ruin". Those odds will also be used as inputs for decisions concerning economic policy or changes in institutional arrangements. To handle all these issues, the first step – but far from the last! – is to represent different instances of price variation by suitable random processes.

The word "suitable" and the plural in "processes" will surprise many readers. It is, indeed, widely believed that "random change" is a synonym for "prices that move up a bit or down a bit following the toss of a coin". The technical term is "simple random walk". It was made popular by a book title that won the high distinction of becoming a cliché, namely, "random walk down the street".

The belief that there is no alternative is strengthened by the fact that coin tossing is, indeed, the oldest and by far the most widely used model of price variation. The (unsaid) point of departure of this book and this survey is that the term "random" has a far broader meaning, allowing the coin tossing model to be replaced by alternatives. Many of the alternatives are "unsuitable", but it will be argued that the alternative I put forward in this book, based on "multifractals", is very suitable indeed.

The multifractal model does not belong to esoteric mathematics and it must not be allowed to remain part of pure science. Its practical consequences are many and very serious. The first but not last is in the spirit of the Hippocratic Oath, "do no harm", which deserves to be generalized to finance and is best expressed in nautical terms. When a ship was built to navigate placid lakes by fair weather, to send it across the ocean in a typhoon season is to do serious harm. Similarly, the "coin tossing" model of financial prices may well be beloved by mathematicians, but it denies the existence of hurricanes; therefore it is dangerous.

The preceding nautical analogy will be heavily used throughout this text, because it resides at the very center of the present study. Not only alternatives to the coin tossing model are available, but the multifractal alternative differs in "qualitative" ways that have immediate consequences for finance and economic policies.

The coin tossing model exemplifies a form of randomness (a "state of randomness", as I shall argue) that can be called "mild". Had the evidence agreed with this model – but it does not at all – variability in finance would be as easily controllable as is variability in physics.

However, the coin tossing model must not be criticized too hard. It is always best to start with the simplest possible model and hold to it until it has begun to bring more harm than value. In its time, it played a fundamental and positive role in creating awareness of the difficulty of even the simplest forms of randomness. One can also argue that for the "man in the street" coin tossing is an adequate description of the facts. But policy makers and the professionals in finance are (or should be) far more demanding. It matters very much for them that, as will be seen, coin tossing is very far from accounting for some essential facts. Once again, the history of price variation is filled with "financial hurricanes" while we shall see

that coin tossing claims that they practically never happen. Ship-builders and ship owners cannot predict the dates and destructiveness of the hurricanes their vessel will encounter over its lifetime. But the knowledge that hurricanes will happen – and realistic evaluation of the corresponding odds – permeates ship-building, ship ownership and navigation.

This book argues that tools needed to acquire some mastery of the intensities of financial hurricanes are already available. They are those of fractal and multifractal geometry, a discipline better known as describing the shapes of coastlines and clouds and the distribution of galaxies and as having led to the discovery of the Mandelbrot set. My claim is that it also describes the growth and collapse of financial prices.

Pick the Fakes

This book includes a multitude of words or numbers and of formulas and dry diagrams. Without mastering them, my claims and contributions cannot be fully understood and appreciated. But to make the central point, as this survey proposes, words, formulas and diagrams are not really necessary.

To explain my central point, the best and quickest way is to encourage the reader to participate in a test concerning figures 1 and 2. No one is asked to accept pictures as the sole or final arbiter in scientific discourse, only as a useful additional tool. Pictures are often used to delude, but in this instance they deserve to be described as uncovering a widespread delusion and assisting in the selection of an improvement.

Among the many graphs in figure 1, some are "real". They follow the practice of financial journals and trace the sequence of daily closing of some price series such as security, commodity, foreign exchange or interest rate. The other graphs in figure 1 will be called "forgeries". They correspond to imitations of financial reality provided by mathematical models that are fully specified in quantitative fashion, therefore can be sampled and illustrated without resorting to unreported stretching and reducing or other such manipulations. Figure 2 plots price differences from one day to next.

Now the "find-the-fakes test" can be described: you are asked to separate reality and forgery as completely as you can. For a perfect score, you must rank the diagrams from "most obviously a forgery" to "apparently real".

When the test relates to figure 1, it is very difficult to separate the real and forged records. Indeed, all such records tend to look alike. This impression is confirmed by looking through the financial press and the books on the mathematics of finance. The optimist will rush to conclude that coin tossing, which is represented by one of the graphs in figure 1, is perfectly

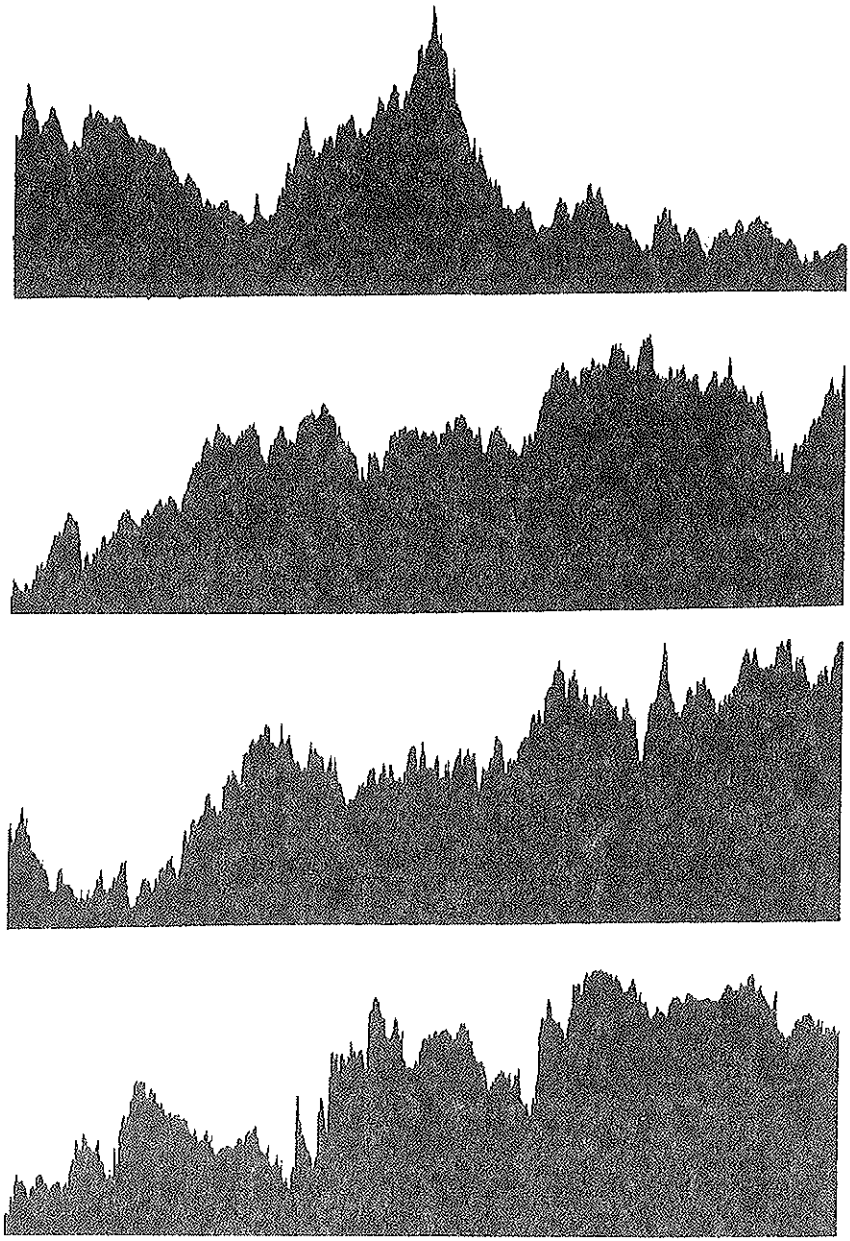


Fig. 1. A collection of diagrams, illustrating – in no particular order – the behavior in time of some actual financial prices and of some mathematical models of this behavior. It would be very difficult to pick the fakes.

acceptable. In fact, we shall see momentarily that this optimism would be seriously misplaced. The resemblance between those curves is due to the fact that graphs of prices themselves do not reveal or enhance important differences, but hide them. In other words, plots of prices are a very inefficient way of presenting information. This is well-known to students of the psychology of vision: position is seen less accurately than change.

In sharp contrast, the lines in the stack of this figure 2 show striking differences between one another. The meaning of those differences will be refined through this survey, ending with the solution of the test.

Large Stock Market Movements and their Odds

Individual investors and professional stock and currency traders know better than ever that prices quoted in any financial market often change with heart-stopping swiftness. Fortunes are made and lost in sudden bursts of activity when the market seems to speed up and the volatility soars. In September 1998, for instance, the stock for Alcatel, a French telecommunications equipment manufacturer, dropped about 40 percent one day and another 6 percent over the next few days. In a reversal, the stock shot up 10 percent on the fourth day. On a longer time scale, most real price changes behave like those in the lower portion of figure 2. However, not all lines at the bottom of on figure 2 are real. (That is, I am not giving away the test the reader is taking!).

The coin tossing model, which served as foundation for the theory of finance used most widely in this century, is represented by the top line of figure 2 (now, I am giving away part of the test). We shall see in a moment that precipitous events like the Alcatel debacle are given totally negligible odds in that theory. Certainly, they should never happen in the lifetime of this generation and the next few. A cornerstone of finance is modern portfolio theory, which tries to maximize returns for a given level of risk. The mathematics underlying portfolio theory ignore the possibility of a typhoon.

This term, coin-tossing, is actually an oversimplification, but the risk-reducing formulas behind portfolio theory rely on a number of demanding premises that are mathematically attractive but rely on hope rather than reality.

First, they suggest that price changes are statistically independent of one another: for example, that today's price has no influence on the changes between the current price and tomorrow's. This is the "efficient market" hypothesis of Louis Bachelier. As a result, prediction of future market movements is never possible.

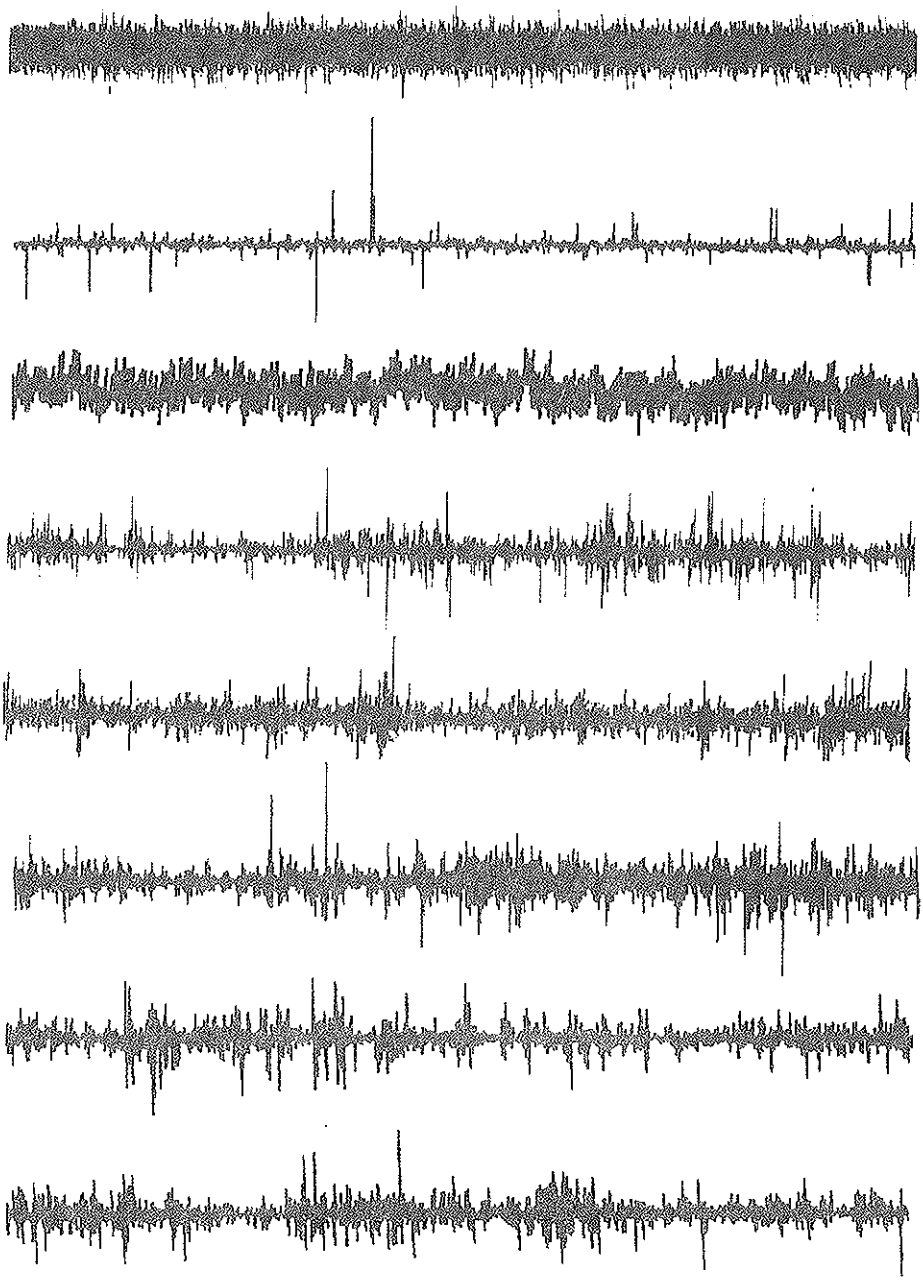


Fig. 2. A stack of diagrams, illustrating the successive "daily" differences in some actual financial prices and some mathematical models. Pick the fakes!

The second assumption is that all price changes are distributed in a pattern that conforms to the standard bell curve. Of the three curves in figure 3, the bell curve is the flattest in the center. The width of its bell (as measured by its "sigma", or standard deviation) depicts how far price changes diverge from the mean. In this perspective, 95% of all cases fall into the narrow range between minus two sigmas and plus two sigmas. As was already mentioned and will be elaborated momentarily, the bell curve declares extreme events to be extremely rare. Typhoons are, in effect, defined out of existence.

Do financial data neatly conform to such assumptions? Of course, they never do! This is shown by a more attentive inspection of the bottom portion of figure 2. It is true that charts of stock or currency changes over time reveal a constant background of small up and down price movements – though not as uniform as one would expect of price changes that fit the bell curve. Invariably, however, these patterns constitute only one aspect of the graph. A substantial number of sudden large changes – spikes on the chart that shoot up and down as with the Alcatel stock – stand out from the background of more moderate perturbations. Moreover, it is typical of the magnitude of price movements – both large and small – to remain roughly constant for an extended period and then suddenly unpredictably increase for another extended period. Big price jumps become more common as the turbulence of the market grows – they cluster on the chart, expressing an obvious amount of dependence.

According to the coin tossing model, these large fluctuations often exceed ten sigmas, meaning ten standard deviations. This value is so huge that standard textbook tables of the Gaussian fail to include it. But a good

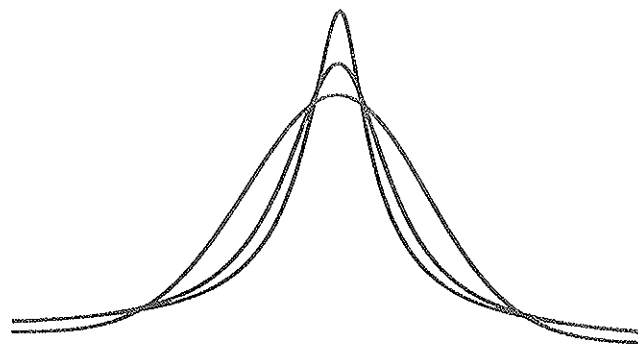


Fig. 3. Shapes of the Gaussian distribution and of two "stable" distributions. The latter provide a far better fit for many financial data, but the multifractal model is even more satisfactory.

calculator should yield their probability a few millionths of a millionth of a millionth of a millionth, that is, approximately one day out of ten million million million years. If risks were so tiny, they would be truly negligible, not worth even a passing thought. But this tiny value grossly contradicts the evidence. The real world of finance produces "ten sigma" spikes on a regular basis – as often as every month, to give an idea – and their probability should be expected to be a few hundredths.

The tiny probability mentioned a few lines above illustrates that the Gaussian practically vanishes near the left and right ends of the graph. Had the horizontal axis been part of figure 3 (which it is not), it would have hidden those insignificant tails.

Reality is far better represented by the other two curves in figure 3, both with more peaked heads and fatter tails. These curves belong to the "M 1963 model" produced by my first attack on financial data, in 1963. Having revealed this fact, it is best to narrow down the test the reader is supposed to be taking. Price changes according to the M 1963 model are the source of the second line in figure 2. This is clearly better than the top line, to be sure, but far from being the last word.

Coin Tossing Normality Versus the Financial Reality

The bell curve is often described as illustrating the "normal" distribution. But does it follow that financial markets should be described as "abnormal or anomalous"? Of course not: they are what they are, and it is the coin tossing model, and therefore the portfolio theory, that are flawed.

Modern portfolio theory poses a danger to those who believe in it too strongly and is a powerful challenge for the theoretician. Though sometimes acknowledging faults in the present body of thinking, the extreme bearish answer is that there is no alternative: large market swings are anomalies, individual "acts of God" that present no conceivable regularity. Other adherents suggest that coin tossing must be maintained, "faute de mieux", because no other premises can be handled through tough mathematical modeling leading to a rigorous quantitative description of at least some features of major financial upheavals.

An increasingly wide agreement is being reached, that the extreme bearish view is untenable and that the coin tossing model must be replaced by one that allows (near-) instantaneous price changes and substantial temporal dependence. This agreement marks a change of mood in the "mainstream", bringing it toward the views I have been campaigning for since 1963 and 1965, respectively. From this point on, however, two general

approaches are in conflict, leading to what I shall call "micromanaged" and "macromanaged" models.

Micromanaged models agree with my diagnosis but not my follow-up. They proceed through a series of "fixes". Each fix "patches" a perceived defect of coin tossing, independently of its other defects. The outcome is that this approach accumulates a large number of parameters and no property is present that was not knowingly incorporated in the construction. In the nautical analog, the fixes consist in lengthening a small boat's keel, lengthening its mast, reinforcing its engine, etc..., one by one.

It is clearly preferable to design a large boat from scratch. Similarly, my experience of successful modeling in other fields has fostered deep *a priori* doubts about the chances of micromanaged modeling in finance. But personal prejudices would not have mattered if *a posteriori* modeling had been effective. I think it is not.

My own work – carried out over many years – takes a very different and decidedly bullish position. I claim that a financial model can be redesigned following an approach that is macromanaged by being guided by a principle of fractal invariance to which we shall come soon. The outcome, as I propose to show, is that the variation of financial prices can be accounted for by a model derived from my work in fractal geometry. Once again, fractals – or their later elaboration, called multifractals – do not claim the ability to predict the future with certainty. But they do create a more realistic picture of market risks. Given the recent troubles confronting the large investment pools called hedge funds, it would be foolhardy not to investigate models providing more accurate estimates of risk.

Fractals, Multifractals and the Market

An extensive mathematical basis already exists for fractals and multifractals. Fractal patterns do not only appear in the price changes of securities but also in the distribution of galaxies throughout the cosmos, in the shape of coastlines and in the decorative designs generated by innumerable computer programs.

A fractal is a geometric shape that can be separated into parts, each of which is a reduced-scale version of the whole. In finance, this concept is not a rootless abstraction but a theoretical reformulation of a down-to-earth bit of market folklore, namely, the notion that movements of a stock or currency all look alike when a market chart is enlarged or reduced so that it fits some prescribed time and price scales. This implies that an observer cannot tell which of the data concern prices that change from week to week, day to day, or hour to hour. This quality defines the charts as fractal

curves and many powerful tools of mathematical and computer analysis become available.

A technical term for this form of close resemblance between the parts and the whole is self-affinity. This property is related to the better-known concept of fractals called self-similarity, in which every feature of a picture is reduced or blown up by the same ratio, a process familiar to anyone who ordered a photographic enlargement or a xerox copy. Financial market charts, however, are far from being self-similar. If we simply focus on a detail of a graph, the features become increasingly higher than they are wide & dash as are the individual up-and-down price ticks of a stock. Hence, the transformation from the whole to the parts must shrink the time scale (the horizontal axis) more than the price scale (the vertical axis). This task can be performed by copiers using lasers. The geometric relation of the whole to its parts is said to be one of self-affinity.

Unchanging properties are not given much weight by most statisticians. But they are beloved of physicists and mathematicians like myself, who call them invariances and are happiest with models that present an attractive invariance property. A good idea of what I mean is provided by drawing a simple chart that inserts (interpolates) price changes from time 0 to a later time 1 in successive steps. The intervals themselves are chosen arbitrarily; they may represent a second, an hour, a day or a year.

The process begins with a price represented by a straight trend line called "initiator", shown in the top panel of figure 4. Next, a broken line called "generator" is used to create the pattern that corresponds to a slow up-and-down price oscillation. It is obviously essential for the number and positions of the pieces of the generator to be completely specified. If it is not, or if one allows oneself the right to fiddle with the generator during the construction, no prediction could be made. In figure 4, the generator consists of three pieces that are inserted (interpolated) along the straight trend line. (A generator with fewer than three pieces could not simulate a price that must be able to move up and down). Then, each of the generator's three pieces is interpolated by three shorter ones. Repeating these steps reproduces the shape of the generator, or price curve, but at increasingly compressed scales. Both the horizontal axis (time scale) and the vertical axis (price scale) are squeezed to fit the horizontal and vertical boundaries of each piece of the generator.

Interpolations Continued (not quite) Forever

Only four construction stages are shown in figure 4, but the same process continues. In theory, it has no end, but in practice, it makes no

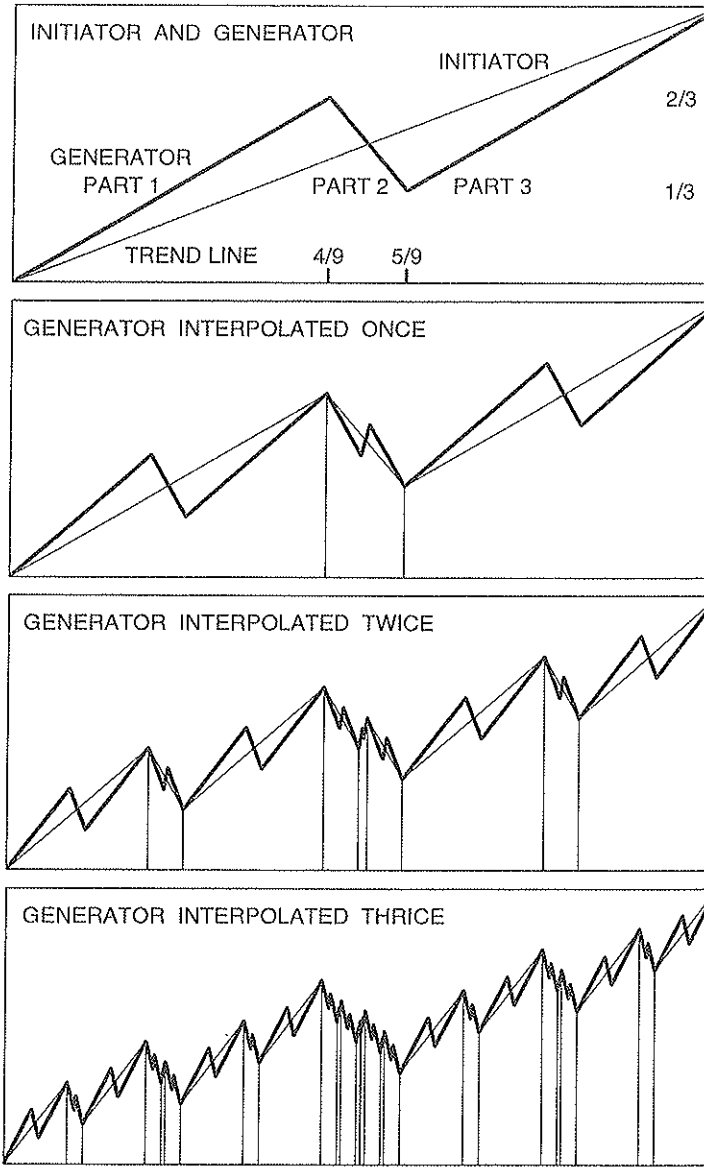


Fig. 4. Constructing a "pseudo-Brownian cartoon" of the idealized coin-tossing model that underlies modern portfolio theory. The construction starts with a linear trend ("the initiator") and breaks it repeatedly by following a prescribed "generator". The interpolated generator is inverted for each descending piece. The pattern that emerges increasingly resembles market price oscillations.

sense to interpolate down to time intervals shorter than those between trading transactions, which may be of the order of a minute. The fact that each piece ends up with a shape like the whole is not a surprise: this scale invariance is present simply because it was built in. The novelty (and surprise) is that these very simply defined self-affine fractal curves exhibit a wealth of structure. The beauty of fractal geometry is that it does not consist in micromanaged models in which everything of interest has been inputed separately. Fractals involve only macromanaged instructions, yet make possible a model general enough to reproduce the patterns that characterize portfolio theory's placid markets as well as the tumultuous trading conditions of real markets. Indeed, the construction's outcome, if plotted as in figure 2, is very sensitive to the exact shape of the generator.

For example, figure 4 uses a very special generator that – according to a theory I developed – will produce a behavior that is pseudo-Brownian, that is, close to the relatively tranquil, "mildly random", picture of the market ruled by coin tossing. But this level of tranquillity prevails only under extraordinarily special conditions that are satisfied only by equally special generators. Figure 3 satisfies those conditions because each generator segment's height – namely, 2/3, 1/3 or 2/3 – was made equal to the square root of the corresponding width – namely, 4/9, 1/9 or 4/9. This "square root rule" is a characteristic of a process physicists call "simple diffusion" Adherence to the assumptions behind this oversimplified model is one of the central mistakes of modern portfolio theory. It is much like a theory of sea waves that forbids the swells to exceed six feet.

A first and very important generalization of figure 4 yields models that are non Brownian but can be called "unifractal". It consists in continuing to require that the height of every segment of the generator be linked to its width by the same relation in the form of a power H. Previously, we set $H = 1/2$, but a different value of H can be chosen, as long as it is positive and less than 1. Taking $H = .7$ suffices to change the top line of figure 2 into its third line. On the corresponding graphs in figure 1, the place of tranquility and mildness is taken by movements that are non-periodic but described by everyone as "cyclic", with many apparent cycle lengths, ranging from very small up to "about three cycles in a sample". (This last rule is a remarkable observation that cannot be elaborated here). Here, cyclic behavior is present in the output without having been incorporated in the input. This is lovely, but large spikes of variation were lost and must be reinstated.

There is a second and far more drastic generalization of figure 3. So far, market activity was assumed constant but one can allow it to speed up and slow down. This variability is the essence of volatility, in fact, practical

people describe the diverse lines at the bottom of figure 2 as proceeding at many different speeds at different times. This is why models that allow for variability add the prefix "multi" before the word "fractal".

To define "activity" is beyond our concern and not necessary. The key idea is that the market does not follow the physical time that procedes with the relentless regularity of a clock, but instead a subjective time that flows slowly during some periods and fast during others.

In this spirit, the theory provides a neat "transmutation" from uni – to multifractal. The key step shown in figure 5 is to lengthen or shorten the horizontal time axis so that the pieces of the generator are either stretched or squeezed. At the same time, the vertical price axis may remain untouched. As seen on the "back" wall of figure 5, the first piece of the unifractal gen-

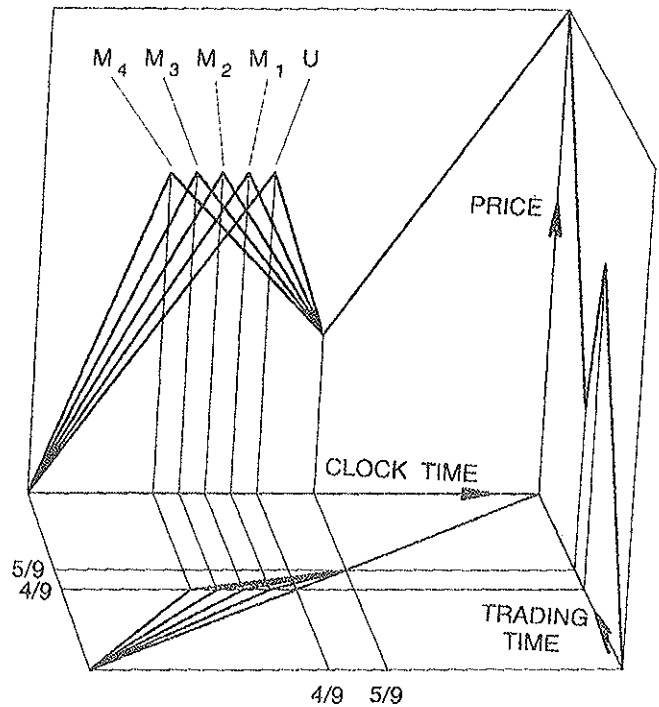


Fig. 5. This open cube illustrates related generators: The "right wall" shows an oscillating generator in trading time. This is the pseudo-Brownian (unifractal) generator of Figure? The "back wall" shows four multifractal oscillating generators in clock time. The "floor" shows the generators that relate the clock time to trading time. Each is an increasing function of the other. Moving a piece of the fractal generator to the left causes the same amount of market activity in a shorter time interval for the first piece of the generator and the same amount in a longer interval for the second piece.

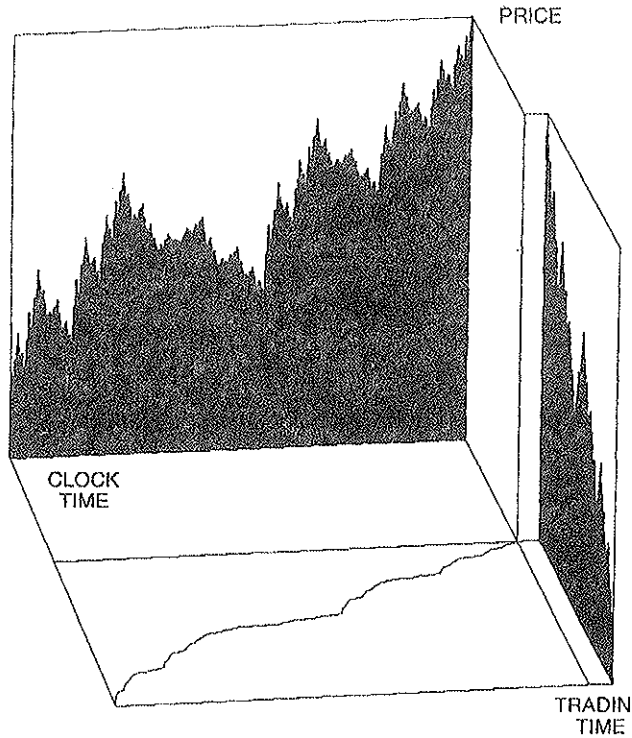


Fig. 6. The underlying pattern is as in figure 5, but limited to the left-most generator, and the generators are replaced by the curve obtained by repeating it recursively as done in figure 3. To make the picture clearer, the back and right wall are moved away from the floor.

erator is progressively shortened, which also provides room to lengthen the second piece. After making these adjustments, the generators become multifractal (M1 to M4). As seen on the "floor", of figure 6, market activity speeds up in the interval of time represented by the first piece of the generator and slows in the interval that corresponds to the second piece.

When the generators in figure 5 are used recursively, one obtains the patterns shown in figure 6. Recall that these patterns do not pretend to exhaust all the possibilities offered by either theory or the facts. Their sole aim is to show the power of the very simplest fractal models. Such an alteration to the generator can produce a full simulation of price fluctuations over a given period, using the process of interpolation described earlier. Each time, the first piece of the generator is further shortened. The process of successive interpolation produces a chart that increasingly resembles the characteristics of volatile markets (figure 7).

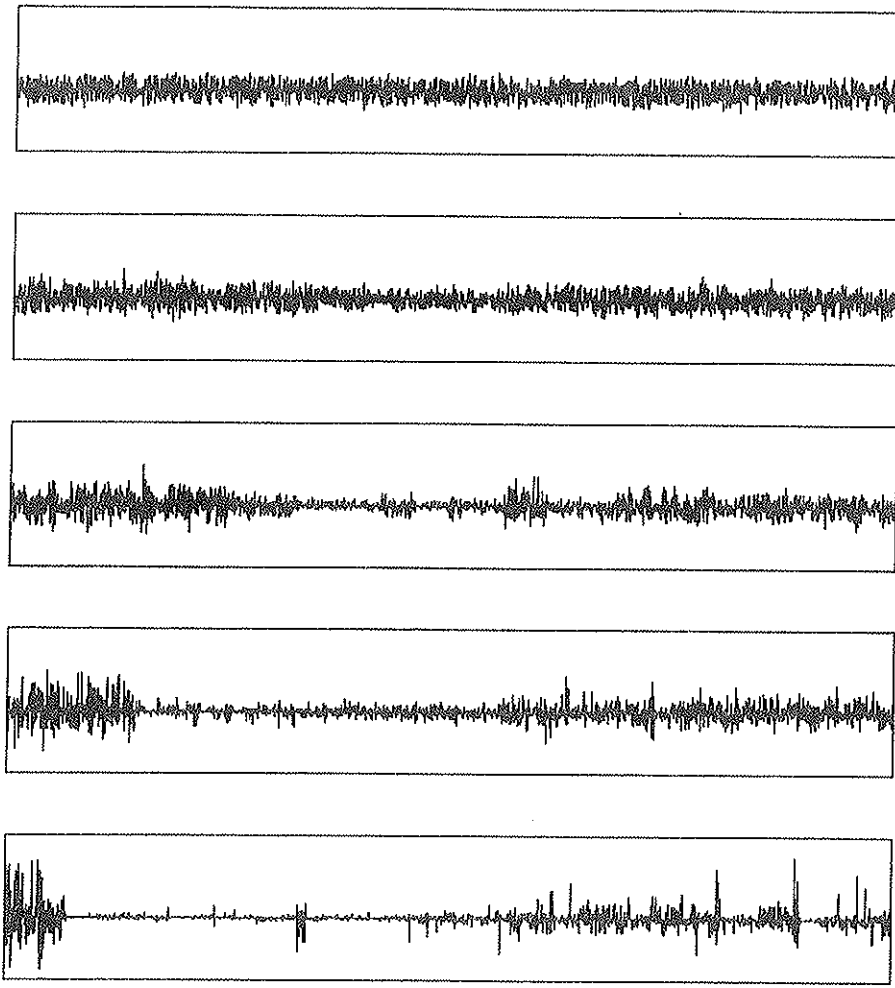


Fig. 7. Randomized multifractal "price increments" that correspond to the five multifractal generators in figure 5, on top a pseudo-Brownian sequence of "price increments". A gradual displacement of the generator to the left causes market activity to increase gradually, becoming more and more volatile.

Once again, the unifractal (U) chart that prevails before any shortening corresponds to the becalmed markets postulated in the portfolio theorists' model. Proceeding down the stack (M1 to M4), each chart diverges further from that model, exhibiting the sharp, spiky price jumps and the persistently large movements that characterize financial trading.

An important detail has not been mentioned yet To make these models of volatile markets achieve the necessary realism, the three pieces of each generator were scrambled – a process not shown in the illustrations. It works as follows. Altogether, the three pieces of the generator allow the following six permutations: 1,2,3; 1,3,2; 2,1,3; 2,3,1; 3,1,2 and 3,2,1.

Conveniently enough, a die has six sides; imagine that each bears the image of one of the six permutations. Before each interpolation, the die is thrown and the permutation that comes up is selected.

Back to the Game of Picking the Fakes

How do simulations of the multifractal model stand up against actual records of changes in financial prices? To respond, let us return to figure 2, which is a composite of several historical series of price changes with a few outputs of artificial models.

As we have already observed, the goal of modeling the patterns of real markets is certainly not fulfilled by the first chart, which is extremely monotonous and reduces to a static background of small price changes, analogous to the static noise from a radio. Volatility stays uniform with no sudden jumps. In a historical record of this kind, daily chapters would vary from one another, but all the monthly chapters would read very much alike.

The rather simple second chart is less unrealistic, because it shows many spikes; however, these are isolated against an unchanging background in which the overall variability of prices remains constant. The third chart has interchanged strengths and failings, because it lacks any precipitous jumps.

The eye tells us that these three diagrams are unrealistically simple. Let us now recall the sources. Chart 1 illustrates price fluctuations in a model introduced in 1900 by French mathematician Louis Bachelier. The changes in prices follow a "random walk" that conforms to the bell curve and illustrates the model that underlies modern portfolio theory. Charts 2 and 3 are partial improvements on Bachelier's work: one is the "M 1963" model I proposed in 1963 (based on Levy stable random processes). The other is the "M 1965" model I published in 1965 (based on fractional Brownian motion). These revisions of coin tossing are inadequate, except under certain special market conditions.

By now, the test around which this survey is structured has been reduced to a careful inspection of the more important five lower diagrams of the graph. Let me now add a piece of information that was withheld until now: at least one record is real and at least one is a computer-generated sample of my latest multifractal model. The reader is free to sort those five lines into the appropriate categories. I hope the forgeries will be perceived as surprisingly effective. In fact, only two are real graphs of market activity. Chart 5 refers to the changes in price of IBM stock and chart 6 shows price fluctuations for the dollar-deutsche mark exchange rate. The remaining charts (4, 7 and 8) bear a strong resemblance to their two real-world predecessors. But they are completely artificial.

Two technical points must be mentioned before moving on to conclusions. The recursive constructions in the body of the paper were nothing but "cartoons". The artificial charts 4, 7 and 8 were, instead, generated through a refined form of my multifractal model, called "fractional Brownian motion in multifractal trading time". Secondly, this introductory Survey necessarily emphasizes graphics, but – once again – the theory of multifractals is endowed with full numerical tools of analysis.

Very Tentative Conclusions: Diversification and Reinsurance

What conclusions should be drawn from all this? Does this matter to a corporate treasurer, currency trader or other market strategists? Does this matter to the central banker and others concerned with overall financial and economic policy? Does this matter to the economist who seeks to explain the workings of the economy and concedes that his task may be helped by an accurate description of part of what is to be explained?

All those questions arise because the discrepancies between coin-tossing and the actual movement of prices have become too obvious to be ignored much longer. Prices do not vary continuously, and they are subjected to wild fluctuations at all time scales. Volatility – far from a static entity to be ignored or easily compensated for – is at the very heart of what goes on in financial markets. In the past, nearly everyone embraced the modern portfolio theory because of the absence of strong alternatives. But one need no longer accept it at face value.

However, the multifractal alternative is very young and very far from being fully developed. It deserves to draw attention (and criticism). By contrast, modern portfolio theory was formulated twenty-four years ago and has been energetically developed ever since. Moreover, wild variability is a new notion endowed with little inherited capital. Modern portfolio theory

inherited a large accumulated capital of techniques that statisticians designed to deal with mild Gaussian variability. The challenge was to adapt them to the context of financial prices.

Therefore, it is necessary, as we near a conclusion, to separate thoughts concerning the near future from thoughts concerning the longer range. Multifractals can immediately be put to work to "stress-test" portfolios, in particular, from the viewpoint of a quantity called "value at risk", whose definition is unfortunately beyond the scope of this text, stress-testing begins by questioning how a portfolio would have performed if it had been designed a while ago. That is, the simplest stress test merely uses historical data. But the actual market test will not come in the past, rather in the future, and a future that simply repeats the past is only one of many alternatives, and not a very likely one. The goal of every model of price variation (coin-tossing not being an exception) is to use the past to create the same patterns of variability as do the unknown rules that govern actual markets. This attempt should yield a collection of alternative scenarios for the future, and Stress-testing should include tests based on many such alternatives.

According to the coin-tossing model, the differences between those alternatives are comparatively slight. Not so with the multifractal models. They describe the past market fluctuations more realistically and the scenarios they propose for the future include a quota of extreme events that will really stress a portfolio. This is all that can be said on this subject at this point.

Of greatest interest, at least to me, are problems that the multifractal model confronts on broader institutional, temporal and spatial horizons. They are more important than any detail and question the worth of the widespread faith in the power of diversification and other forms of lumping and averaging. Here an enlightening analogy and powerful guidance for the future is provided by a distinction between different levels of insurance that relates to my distinction between mild and wild "states" of chance.

Most life, automobile, or homeowner risks are mild. Very much like the coin-tossing model of price change, they fall within a narrow range and are mutually independent. Even when a portfolio of insurance contracts is small, the wonders of diversification (due to the law of large numbers and related mathematics) can be trusted to create a risk of ruin that is sufficiently small to be profitable even for an insurance company with limited reserves. To play safe and to insure the occasional higher risk, the insurer of mild risks will seek reinsurance – which will seldom be needed, therefore will not be expensive. When a tornado defeats diversification of homeowner policies, the reinsuer is likely to be an entity that collected no premiums, namely an agency of a state.

However, many other risks seeking to be insured are wild, very much like

in my multifractal model of price change. They involve the equivalents of the notorious "ten sigma" price changes that were discussed earlier in this survey. Ordinary diversification would be defeated by such risks, even if the number of cases had sufficed for a law of large numbers to apply. More precisely, the odds of those wild risks, if included in the usual calculations, would imply reserves that are clearly unreasonable. However, such risks become insurable if they are immediately shared with reinsurers (or almost directly with competitors, as is apparently the case in the shipping industry).

The key fact is that insurers cannot survive by only considering the "fair weather" 95% of the claims, which would have easily been diversified. Not only the 5% of large "foul weather" claims cannot be ignored, but their odds are non-negligible and are an essential input of planned and carefully priced reinsurance.

Once again, theories based on coin-tossing legislate this "foul weather" out of existence, but it is evident that many features of the real world are best understood as designed to tackle comparatively rare but potentially disastrous situations. It is indeed filled with state or private institutions and informal or ad-hoc arrangements whose purpose can be viewed as that of reinsurance. A central part of my thinking in finance is that those arrangements may have worked in the past but cannot be relied upon in the future. As to institutions, their role deserves a fresh examination.

As a result of globalization, the relevance of the preceding comments on insurance is bound to increase. Under the coin-tossing model, the effects of globalization are limited. But the actual behavior of financial prices confirms what intuition suggests: the larger the markets, the greater the attention demanded by the potentially disastrous effects of financial storms.

To conclude, no overall mathematical technique comes close to forecasting a price drop or rise on a specific day on the basis of past records. Multifractals do not do any better. But they provide estimates of the probability of what the market might do in the future and allow one to prepare for inevitable sea changes. The new modeling techniques are designed to cast a light of order into the seemingly impenetrable thicket of the financial markets. They also recognize the mariner's warning that, as recent events demonstrate, deserves to be heeded: on even the calmest sea, a gale may be just over the horizon.

REFERENCES

Mandelbrot, B., *Fractals and Scaling in Finance: Discontinuity, Concentration, Risk* (Springer-Verlag, 1997).

Mandelbrot, B., *Multifractals and 1/f Noise: Wild Self-Affinity in Physics* (Springer-Verlag, 1999).

Mandelbrot, B., Calvet, L., and Fisher, A., 'The Multifractal Model of Asset Returns', three reports of the Cowles Foundation for Economic Research of Yale University. Web address of the first: http://papers.ssrn.com/sol3/paper.taf?/ABSTRACT_ID=78588. Addresses of the other two: same beginning and the endings _ID=78608 and _ID=78628.

COOPERATIVE BEHAVIOR IN SIMPLE AND COMPLEX

JOEL L. LEBOWITZ

I will first give an overview of how statistical mechanics works for simple systems and then discuss some successes and failures in the attempts to extend it to cooperative behavior in complex systems. I will also discuss the question of the social responsibility of scientists *qua* scientists.

Nature has a layered hierarchical structure, with time, length and energy scales ranging from the submicroscopic to the supergalactic. A central lesson of the very successful advances in the physical sciences during the past three hundred years is that while there are many new phenomena as one proceeds from the consideration of individual entities to that of aggregates, there are no new fundamental laws, and explanations always go from smaller to larger scales. (It should also be noted, however, that some of the mathematical dualities currently being studied in string theory suggest possible direct connections between the highest and the lowest level, closing the loop).

On the other hand there are higher level phenomena, including some of the most interesting ones to us humans, e.g. life, whose origin in the more fundamental lower level laws is as yet little understood. Even where the connection between the levels is more transparent, the specific form of complex higher level phenomena cannot in general be deduced, in any meaningful way, directly from the behavior of its microscopic constituents; think of the weather and the motion of the air molecules. In fact, due to the scale separation between different levels, it is often possible and sometimes essential to discuss different levels independently – quarks are really irrelevant for biological processes and atoms are a distraction when studying ocean currents. Still, deeper understanding always results when the cooperative behavior of an aggregate can be traced back to properties of its constituents. Such understanding is often a prerequisite for innovation, as in the design of new materials or of new drugs.

Statistical mechanics provides a framework for describing how well-defined higher level patterns of organized behavior may result from the activity of a multitude of interacting lower level individual entities. The subject was developed for, and has had its greatest success so far in, relating macroscopic thermal phenomena to the microscopic dynamics of atoms and molecules. Some of these phenomena can be understood as the additive effects of the actions of individual atoms, while others are paradigms of emergent cooperative behavior, having no direct counterpart in the properties or dynamics of the microscopic constituents considered in isolation.

An example of the former is the pressure exerted by a gas on its container; this is just the sum of impulses due to atoms colliding with the walls. A particularly fascinating example of the latter is the observed irreversible behavior of macroscopic systems, as in the tendency of an isolated macroscopic system to evolve towards equilibrium – a state characterized by the maximization of the entropy function under the relevant constraints. The explanation of why and how such time asymmetric collective behavior arises from completely reversible microscopic dynamics is one of the great achievements of the founders of statistical mechanics, J.C. Maxwell, W. Thomson and L. Boltzmann. Other examples of emergent phenomena, well explained by statistical mechanics, are phase transitions in equilibrium systems, such as occur in the boiling or freezing of a liquid. Here dramatic changes in structure and behavior of the macroscopic system are brought about by small changes in the temperature or pressure without any change in the individual atoms or molecules making up the fluid. (The volume occupied by a kilo of water molecules at atmospheric pressure in the Vatican, while changing only very little when the temperature increase is between $5^o C$ and $95^o C$, increases by a very large factor when its temperature changes from $99.9999^o C$ to $100.0001^o C$).

The methods of statistical mechanics used to understand and predict these phenomena owe their success to the facts that, (1) due to the separation of energy scales alluded to earlier, atoms or molecules (or possibly ions and electrons) can be taken as the basic building blocks of matter at temperatures below about one million $^o C$, (above which nuclear structure may become relevant), and (2) that even very crude modeling of these atoms and molecules yields many essential features of their collective behavior.

Statistical mechanics therefore often takes as its lowest level starting point Feynman's description of atoms as "little particles that move around in perpetual motion, attracting each other when they are a little distance apart, but repelling upon being squeezed into one another". Why this crude classical picture (a refined version of that held by some ancient Greek philosophers) gives predictions about the properties of many sub-

stances which are not only qualitatively correct but in many cases also quantitatively accurate, is not entirely clear.

An important question is whether and how we might adapt the methods of statistical mechanics, which may be characterized as "subdivide", "simplify", and "average" (all up to a certain point), to higher level collective systems in which the relevant basic constituents are themselves complex entities. Here the interest is not so much in typical equilibrium behavior, or even in the approach to equilibrium, as in the dynamic responses of open systems, like the biosphere, evolving over long times in an environment subjected to a variety of regular and irregular (sometimes best modeled as random) influences. What may be particularly important in some of these systems are rare but powerful events (avalanches, catastrophies) which have long term effects. Our understanding of such far from equilibrium systems is at present unfortunately very poor even for simple atomic systems, such as the turbulent flow of water in a pipe. It still lags far behind that of equilibrium systems, the approach to equilibrium or stable stationary non-equilibrium states.

There have nevertheless been many attempts to apply statistical mechanical principles to complex systems. Some of these have achieved a certain degree of success; examples include the self sorting of different cells in an organism, the flow of vehicular traffic and the formation of galaxies. This suggests that internal complexity and individuality need not exclude cooperative behavior amenable to statistical mechanical analysis. There is however something special about each of these systems. For example, the reason that statistical mechanical methods work well when we treat stars as point particles is clearly due to the fact that the aspect of their behavior in which we are interested, e.g. their tendency to aggregate into clusters and galaxies, depends on a property, gravitational attraction, with respect to which their internal structure and individual differences play little or no role. Similar considerations apply to the use of statistical mechanical methods for the behavior of vehicular traffic, where transitions between "free flow" and "jam" traffic can, to a certain extent, be treated as phase transitions. Here the special nature of the situation comes from the fact that it deals with structured and highly restricted possibilities of behavior.

Still the above examples show that applications of statistical mechanical methods to certain types of complex systems can indeed yield useful insights. The studies of traffic flow have apparently been successful in predicting patterns and suggesting actions for better traffic management. This leads us to hope that we might be able to apply the methods of statistical mechanics, or at least the more general methods of theoretical physics, also to the analysis of those complex collective systems whose behavior is of

direct relevance to our survival and well being. Perhaps the most important such system is human society itself.

The goals of such studies go beyond the applications of science to the development of useful technologies. The type of analysis I have in mind would ideally provide us with insights needed to guide technology towards achieving basic humanistic goals, ranging from improvements of the environment to the elimination of hunger and social injustice. Such guidance of technology and applied science in general is clearly desirable. For while many practical applications of science have greatly advanced the quality of the lives of most people – witness the green revolution and the improvements in health together with increased longevity – science can be, often has been been, and unfortunately continues to be used also for evil purposes. While technology is indeed very important for improving the human condition, by itself it is far from sufficient.

What we scientists therefore need to do within our discipline is to develop a moral vision inspired by science itself. This is the vision of an unimaginably grand universe of which living beings and particularly conscious ones, like us humans, are a very small but unique component – a component able to appreciate and formulate general fundamental laws, unifying the great variety of existing levels. So while I agree with Einstein that, "scientific statements of facts and relations cannot produce ethical statements", it is also my belief – and was Einstein's – that the scientific perspective strongly resonates with humanistic values. This perspective makes differences between peoples based on nationality, race, belief or gender entirely trivial, while making the things humans have in common, e.g. the ability to comprehend many aspects of our universe, very special and significant. Indeed, the scientific outlook should make us scientists evangelists for a sustainable and just world.

I very much hope that this meeting will contribute to making scientists aware of their special responsibility towards all humans and indeed to all life – a responsibility which arises from our ability to appreciate the grand and subtle structure of the universe we inhabit.

Acknowledgments

I thank S. Goldstein and E. Speer for their very helpful comments.

REFERENCES

Bak, P., *How Nature Works, The Science of Self-Organized Criticality* (Springer Verlag, 1996).

Dyson, F.J., *The Sun, the Genome and the Internet: Tools of Scientific Revolution* (Oxford University Press, April 1999).

Lebowitz, J.L., 'Statistical Mechanics: A Selective Review of Two Central Issues', *Reviews of Modern Physics*, 71, 1999, pp. 346-357.

Lebowitz, J.L., 'Boltzmann's Entropy and Time's Arrow', *Physics Today*, 46, 1993, pp. 32-38,

Mandelbrot, B., *The Fractal Geometry of Nature* (Freeman, 1983).

Wolf, D.E., Schreckenberg, M., and Bachem, A. (eds.), *Traffic and Granular Flow* (World Scientific, Singapore, 1996).

EXTREME DEVIATIONS AND APPLICATIONS

URIEL FRISCH and DIDIER SORNETTE

1. Introduction

Consider the sum

$$S_n \equiv \sum_{i=1}^{n} x_i \,, \tag{1}$$

where the x_i are independent identically distributed (iid) random variables with probability density function (pdf) $p(x)$ and mean value $\langle x \rangle$. The central limit theorem ensures, with suitable conditions such as the existence of finite second-order moments or refinements thereof [1, 2], that as $n \to \infty$ the pdf of $\frac{S_n - n\langle x \rangle}{\sqrt{n}}$ becomes Gaussian. In other words, the "typical" fluctuations of S_n/n around its mean value are Gaussian and $O(1/\sqrt{n})$. The large deviations theory is concerned with events of much lower probability when S_n/n deviates from its mean value by a quantity $O(1)$ [3-6]. In non-technical terms, for large n,

$$\text{Prob}\left[\frac{S_n}{n} \simeq y\right] \sim \mathrm{e}^{n s(y)} \,, \tag{2}$$

where $s(y) \leq 0$ is the Cramér function (also called "rate function").

In this paper we are concerned with "extreme deviations", that is the régime of *finite n* and *large* $S_n = X$. This régime exists only when the pdf extends to arbitrary large values of x (*i.e.* has noncompact support). In other words, we are interested in tail behavior. While it is common in statistics to consider test probabilities of the order of 1%, much smaller probabilities are of interest in many areas in which crisis may ensue. If, for instance, one wishes to investigate whether a chemical substance causes cancer, one will be interested in very small test probabilities to make a con-

vincing case. In the field of reliability, failure and rupture, for instance of industrial plants, very small probabilities are the rule. Examples are the calculation of the probability of a defect item passing an inspection system and the calculation of the reliability of a system. 10^{-6} is the probability threshold beyond which the U.S. Food & Drug administration considers that any risk from a food additive is considered too small to be of concern. In the same spirit, the legal U.S. maximum man-made risk to public is 5×10^{-5}.

The main result about extreme deviations for sums of random variables is presented in section 2. The relation to large deviations theory and to other work on extreme deviations is briefly discussed in section 3. Multiplication of random variables is considered in section 4. Applications are presented in section 5. The appendix is devoted to a rigorous derivation of the main result for sums of independent variables.

2. Extreme Deviations for Sums of Random Variables

We are interested in the tail behavior for large arguments x of sums of iid random variables x_i, $i = 1, 2, \ldots$ We shall only consider large positive values of x. All the results can be adapted *mutatis mutandis* to large negative values. We assume that the common probability distribution of the x_i's has a pdf, denoted $p(x)$, which is normalized:

$$\int_{-\infty}^{\infty} p(x)\, dx = 1 , \tag{3}$$

and which can be represented as an exponential:

$$p(x) = e^{-f(x)} , \tag{4}$$

where $f(x)$ is indefinitely differentiable.[1] We rule out the case where $f(x)$ becomes infinite at finite x; this would correspond to a distribution with compact support which has no extreme deviations.

The key assumptions are now listed. All statements involving a limit are understood to be for $x \to +\infty$.

[1] As we shall see in the appendix, this assumption can be relaxed.

(i) $f(x) \to +\infty$ sufficiently fast to ensure the normalization (3);

(ii) $f''(x) > 0$ (convexity), where f'' is the second derivative of f;[2]

(iii) $\lim \dfrac{f^{(k)}(x)}{(f''(x))^{k/2}} = 0$, for $k \geq 3$, where $f^{(k)}$ is the kth derivative of f. The $k = 3$ instance of (iii) will be denoted by $\widetilde{\text{(iii)}}$.

An important consequence of $\widetilde{\text{(iii)}}$ is

$$\lim x^2 f''(x) = +\infty , \tag{5}$$

the proof of which is given in the appendix (Lemma 1).

We introduce now the pdf $P_n(x)$ of $S_n = \sum_1^n x_i$ which may be written as a multiple convolution:

$$P_n(x) = \underbrace{\int \cdots \int}_{n} e^{-\sum_{i=1}^n f(x_i)} \, \delta\left(x - \sum_{i=1}^n x_i\right) dx_1 \cdots dx_n . \tag{6}$$

All integrals are from $-\infty$ to $+\infty$. The delta function expresses the constraint on the sum.

We shall show that, under assumptions (i)-(iii), the leading-order expansion of $P_n(x)$ for large x and finite $n \geq 1$ is given by

$$P_n(x) \simeq e^{-nf(x/n)} \frac{1}{\sqrt{n}} \left(\frac{2\pi}{f''(x/n)}\right)^{\frac{n-1}{2}} , \text{ for } x \to \infty \text{ and } n \text{ finite.} \tag{7}$$

Furthermore, we shall show that the leading contribution comes from individual terms in the sum which are *democratically localized*. By this we understand that the conditional probability of the x_i's, given that the sum is x, is localized, for large x, near

$$x_1 \simeq x_2 \cdots \simeq x_n \simeq \frac{x}{n} . \tag{8}$$

In this section we give a derivation of this result using a formal asymptotic expansion closely related to Laplace's method for the asymptotic evaluation of certain integrals [8].[3] In the appendix we shall give a rigorous proof.

---

[2] This is called log-concavity of the density by Jensen [7] whose work we shall comment on in section 3.

[3] This method is sometimes referred to as "steepest descent", an inadequate terminology when $f(x)$ is not analytic.

To evaluate (6) for $n \geq 2$, we define new variables

$$h_i \equiv x_i - \frac{x}{n}, \quad \text{for } i = 1, \ldots, n-1 \tag{9}$$

$$h_n \equiv -(h_1 + \ldots + h_{n-1}), \tag{10}$$

and the function

$$g_n(x; h_1, \ldots, h_{n-1}) \equiv \sum_{i=1}^{n} f\left(\frac{x}{n} + h_i\right). \tag{11}$$

We can then rewrite (6) as

$$P_n(x) = \underbrace{\int \cdots \int}_{n-1} e^{-g_n(x; h_1, \ldots, h_{n-1})} \, dh_1 \ldots dh_{n-1}. \tag{12}$$

The function g_n has the following Taylor expansion in powers of the h_i's:

$$g_n = nf\left(\frac{x}{n}\right) + \frac{1}{2!} f''\left(\frac{x}{n}\right) \sum_{i=1}^{n} h_i^2 + \frac{1}{3!} f'''\left(\frac{x}{n}\right) \sum_{i=1}^{n} h_i^3 + \cdots. \tag{13}$$

Note the absence of the term linear in the h_i's since, by (10), $\sum_{i=1}^{n} h_i = 0$.

If we momentarily ignore the terms of order higher than two in (13), we obtain for $P_n(x)$ a Gaussian integral the convergence of which is ensured by the convexity condition (ii). This integral is evaluated by setting $y = 0$ and $\lambda = (1/2)f''(x/n)$ in the identity.[4]

$$\underbrace{\int \cdots \int}_{n-1} e^{-\lambda[h_1^2 + \cdots + h_{n-1}^2 + (y - h_1 - \cdots - h_{n-1})^2]} \, dh_1 \cdots dh_{n-1} = \frac{1}{\sqrt{n}} \left(\frac{\pi}{\lambda}\right)^{\frac{n-1}{2}} e^{-\frac{\lambda}{n} y^2}. \tag{14}$$

We thereby obtain the desired asymptotic expression (7) for $P_n(x)$.

We now show that higher than second order terms in the taylor expansion (13) do not contribute to the leading-order result. The quadratic form $(1/2)f''(x/n) \sum_{i=1}^{n} h_i^2$ in the $n-1$ variables $h_1, \ldots, h_{n-1}$ can be diagonalized (it is just proportional to the square of the Euclidean norm in the subspace

[4] This identity if obtained, after proper normalization, by evaluating the n-fold convolution of a Gaussian distribution of variance $1/(2\lambda)$ with itself, which is a Gaussian of variance $n/(2\lambda)$.

$\sum_{i=1}^{n} h_i = 0$). One can show by recurrence that it has $n-2$ eigenvalues equal to $(1/2)f''(x/n)$ and one eigenvalue n times larger. Hence, the Gaussian multiple integral comes from h_i's which are all $O(1/\sqrt{f''(x/n)})$ or smaller. For such h_i's, it follows from the assumption (iii) that all higher order terms are negligible for large x. Furthermore, the scatter of the x_i's around the value x/n, measured by the the rms value of the h_i's is $O(1/\sqrt{f''(x/n)})$. By (5), this is small compared to x, which proves the *democratic localization* property (8).

We shall also make use of a weaker result obtained by taking the logarithm of (7), namely

$$\ln P_n(x) \simeq -nf(x/n), \quad \text{for } x \to \infty \text{ and } n \text{ finite.} \tag{15}$$

This weaker form holds only if

$$\frac{\ln f''(x/n)}{f(x/n)} \to 0\,. \tag{16}$$

We make a few remarks. Our derivation is reminiscent of the derivation of Laplace's asymptotic formula for integrals of the form $\int e^{-\lambda f(x)}\,dx$ when $\lambda \to \infty$, as given, *e.g.*, in reference [8]. The main difference is that in Laplace's method, when f is Taylor expanded around its minimum, terms of order higher than two give contribution smaller by higher and higher inverse powers of λ, so that a single small parameter $1/\lambda$ is enough to justify the expansion, whereas here we made an infinite number of assumptions (iii) for all $n \gtrsim 3$. Actually, it will be shown in the appendix that the sole assumption (iii) with a slight strengthening of (5) is enough to derive the leading-order term (7).

It is easily checked that our result is not equivalent to the well-known fact that the most probable increment $\Delta x/\Delta t$ of a random walk conditioned to go from (x, t) to (x', t') is constant and equal to the average slope $(x' - x) / (t' - t)$; in other words, the most probable path is then a straight line, corresponding to a constant reduced running sum.

The convexity of $f(x)$ at large x is essential for our result to hold. For instance, pdf's with powerlaw tails $p(x) \propto x^{-(1+\mu)}$ give $f(x) = (1+\mu) \ln x$ which is concave. The extreme deviations of the sum S_n are then controlled by realizations where just one term in the sum dominates. This extends to *arbitrary* exponents μ, in this extreme deviations régime, the well-known result that the breakdown of the central limit theorem for $\mu < 2$ stems from the dominance of a few large terms in the sum. The breakdown of demo-

cratic localization far in the tail also happens for pdf's with finite moments of all orders, for example, when $p(x) \propto x^{-\ln x}$ at large x. Here, again the function $f(x) = \ln^2 x$ is not convex.

The result (15) can be formally[5] generalized to the case of dependent variables with nonseparable pdf's $p(x_1, x_2, \ldots, x_i, \ldots, x_n) = \exp[-f(x_1, x_2, \ldots, x_i, \ldots, x_n)]$ where $f(x_1, x_2, \ldots, x_i, \ldots, x_n)$ is symmetric and convex. Indeed, $\frac{\partial f}{\partial x_i}\Big|_{x_1=x_2=\ldots=x_n=S_n/n}$ is then independent of i and the matrix of second derivatives $\partial^2 f / \partial x_i^2$ evaluated at $x_1 = x_2 = \ldots = x_n = S_n/n$ is positive, ensuring that f is minimum at $x_1 = x_2 = \ldots = x_n = S_n/n$ thereby providing the major contribution to the convolution integral.

3. Relation with the Theory of Large Deviations

We now assume, in addition to conditions (i)-(iii) of section 2, that the characteristic function

$$Z(\beta) \equiv \langle e^{-\beta x} \rangle = \int_{-\infty}^{\infty} e^{-\beta x} p(x)\, dx \tag{17}$$

exists for all real β's (Cramér condition). Recall that the Cramér function $s(y)$ is determined by the following set of equations (see, *e.g.*, Refs. [4, 6, 9]):

$$s(y) = \ln Z(\beta) + \beta y\,, \tag{18}$$

$$\frac{ds(y)}{dy} = \beta\,. \tag{19}$$

Hence, $s(y)$ is the Legendre transform of $\ln Z(\beta)$.

Comparison of (2) with (15) shows that the Cramér function $s(y)$ becomes equal to $-f(y)$ for large y. We can verify this statement by inserting the form $p(x) = e^{-f(x)}$ into (17) to get $Z(\beta) \sim \int_{-\infty}^{\infty} dx\, e^{-\beta x - f(x)}$. For large $|\beta|$, we can then approximate this integral by Laplace's method, yielding $Z(\beta) \sim e^{-\min_x(\beta x + f(x))}$. Taking the logarithm and a Legendre transform, we recover the identification that $s(y) \to -f(y)$ for large y. Laplace's method is justified by the fact that $|y| \to \infty$ corresponds, in the Legendre transformation, to $|\beta| \to \infty$. A number of more precise results are known, which relate the tail probabilities of random variables to the large-y behavior of the Cramér function.

---

[5] Additional assumptions are then needed to make sure that higher than second-order terms in the Taylor expansion are not contributing.

For example, Broniatowski and Fuchs [10] give conditions for the asymptotic equivalence of $s(y)$ (called by them the "Chernov function") and of $-\ln \bar{F}(y)$ where $\bar{F}(y) \equiv \int_y^\infty p(x)\,dx$.

A consequence is that the large and extreme deviations régime overlap when taking the two limits $n \to \infty$ and $S_n/n \to \infty$. Indeed, large deviations theory usually takes $n \to \infty$ while keeping S_n/n finite, whereas our extreme deviations theory takes n finite with $S_n \to \infty$. Our analysis shows that, in the latter régime, Cramér's result already holds for finite n. The true small parameter of the large deviations theory is thus not $1/n$ but $\min(1/n, n/S_n)$.

A paper by Borokov and Mogulskii [11] contains a result resembling somewhat ours. Their equation (12) of section 1 states, in our notation, that

$$s_n(y) = ns(y/n)\,, \tag{20}$$

where $s_n(y)$ is the Cramér function for the sum of n independent and identically distributed copies of a random variable with Cramér function $s(y)$. If we identify the tail of the Cramér function with minus the logarithm of the (tail of the) pdf, their result becomes identical with (15). However, their result makes no use of the convexity assumption without which our result will generally not hold.

Broniatowski and Fuchs [10] derive a more general but weaker theorem on the cumulative distribution of S_n for finite n, which ressembles somewhat our result on the *democratic localization* property (8). It is more general because it is valid for pdf's not obeying the convexity condition (5), for example, the Cauchy distribution. It is weaker because it states only that there is a number $\alpha_n > 0$ such that

$$\ln \text{Prob}\,(S_n \geq nx) = \alpha_n[1 + o(1)]\ln \text{Prob}\,(\min(x_1,\ldots x_n)) \geq x)\,, \tag{21}$$

for $x \to \infty$. Roughly speaking, (21) means that the main contributions to the event $S_n \geq nx$ come from the realizations where *all* variables constituting the sum are larger than x, a much weaker statement than the property of *democratic localization* (8).

Jensen [7] also considers the case where n is finite and the tail probability tends to zero, for particular choices of the pdf. Jensen is able to show in a few examples that, even though there is no asymptotics, *i.e.* there is no n tending to infinity, the saddlepoint expansion allows one to get the correct order of the probabilities in the tail, using the so-called tilted density introduced by Esscher [12]. Coupled with the Edgeworth expansion, this leads to results similar to ours. Our work generalizes and systematizes these partial results by providing general conditions of applications, in particular not requiring that f be Taylor expandable to all orders (see the appendix).

4. Multiplications of Random Variables

Consider the product

$$X_n = m_1 m_2 \cdots m_n \tag{22}$$

of n independent identically distributed positive[6] random variables with pdf $p(x)$. Taking the logarithm of X_n, it is clear that we recover the previous problem (1) with the correspondence $x_i \equiv \ln m_i$, $S_n \equiv \ln X_n$ and $-f(x) = \ln p(e^x) + x$. Assuming again the set of conditions (i), (ii) and (iii) on f, we can apply the extreme deviations result (15) which translates into the following form for the pdf $P_n(X)$ of X_n at large X:

$$P_n(X) \sim [p(X^{1/n})]^n, \qquad \text{for } X \to \infty \text{ and } n \text{ finite.} \tag{23}$$

(In this section we omit prefactors; this amounts to using (15) instead of (7)). Equation (23) has a very intuitive interpretation: the tail of $P_n(X)$ is controlled by the realizations where all terms in the product are of the same order; therefore $P_n(X)$ is, to leading order, just the product of the n pdf's, each of their arguments being equal to the common value $X^{1/n}$.

When $p(x)$ is an exponential, a Gaussian or, more generally, of the form $\propto \exp(-Cx^\gamma)$ with $\gamma > 0$, then (23) leads to stretched exponentials for large n. For example, when $p(x) \propto \exp(-Cx^2)$, then $P_n(X)$ has a tail $\propto \exp(-CnX^{2/n})$.

Note that (23) can be obtained directly by recurrence. Starting from $X_{n+1} = X_n x_{n+1}$, we write the equation for the pdf of X_{n+1} in terms of the pdf's of x_{n+1} and X_n:

$$\begin{aligned} P_{n+1}(X_{n+1}) &= \int_0^\infty \mathrm{d}X_n P_n(X_n) \int_0^\infty \mathrm{d}x_{n+1} p(x_{n+1}) \delta(X_{n+1} - X_n x_{n+1}) \\ &= \int_0^\infty \frac{\mathrm{d}X_n}{X_n} P_n(X_n) p\left(\frac{X_{n+1}}{X_n}\right). \end{aligned} \tag{24}$$

The maximum of the integrand occurs for $X_n = (X_{n+1})^{(n+1)/n}$ at which $X_n^{1/n} = X_{n+1}/X_n$. Assuming that $P_n(X_n)$ is of the form (23), the formal application of Laplace's method to (24) then directly gives that $P_{n+1}(X_{n+1})$ is of the same form.[7] Thus, the property (23) holds for all n to leading order in X.

---

[6] What follows is immediately extended to the case of signed m_i's with a symmetric distribution.

[7] Control over higher-order terms in the asymptotic expansion requires, of course, the same conditions (i)-(iii) as in section 2.

Some generalizations are easily obtained. For instance, for exponential distributions, we can allow for different characteristic scales α_j defined by $p_j(x) = \alpha_j e^{-\alpha_j x_j}$. Equation (23) then becomes

$$P_n(X) \sim \exp\left(-n\left[X\prod_{j=1}^{n}\alpha_j\right]^{1/n}\right) \quad \text{for} \quad X_n > \prod_{j=1}^{n}\frac{1}{\alpha_j}. \tag{25}$$

Similarly, if $p_j(x) = \dfrac{2}{\sqrt{2\pi}\sigma_j}\, e^{-x_j^2/2\sigma_j^2}$, with $x_j \geq 0$, we obtain

$$P_n(X) \sim \exp\left(-\frac{n}{2}\left[\frac{X^2}{\prod_{j=1}^{n}\sigma_j^2}\right]^{\frac{1}{n}}\right) \quad \text{for} \quad X_n > \prod_{j=1}^{n}\sigma_j. \tag{26}$$

5. Applications

Considering the simplicity and robustness of the results derived above, we expect the extreme deviation mechanism to be at work in a number of physical or other systems. We are thinking in particular of the application of our result to simple multiplicative processes, that might constitute zeroth-order descriptions of a large variety of physical systems, exhibing anomalous pdf and relaxation behaviors. There is no generally accepted mechanism for their existence and their origin is still the subject of intense investigation. The extreme deviations régime may provide a very general and essentially *model-independent* mechanism, based on the extreme deviations of product of random variables.

Fragments are often found to be distributed according to power law distributions [13]: In section 5.1, we propose a multiplicative fragmentation model in which the exponent is controlled by the depth of the cascading process. Anomalous relaxations in glasses have been largely documented to occur according to stretched exponentials [14, 15]. In section 5.2, we construct a relaxation model based on the idea that a complex disordered system can be divided into an ensemble of local configurations, each of them hierarchically ordered. Stretched exponential pdf are observed in turbulent flow (see, *e.g.*, ref. [9]) and our extreme deviation theory provides a simple scenario (sect. 5.3). Let us finally mention the question of stock market prices and their distribution. Here, the very nature of the pdf's is still debated [16, 17]. While price variations at short time scales (minutes to hours) are well-fitted by truncated Lévy laws [18], other alternative have been proposed [16]. We have found that a stretched exponential pdf pro-

vides an economical and accurate fit to the full range of currency price variations at the daily intermediate time scale. We will come back in future work to document this claim and to describe the relevance of the multiplicative processes studied here.

5.1. *Fragmentation*

Fragmentation occurs in a wide variety of physical phenomena from geophysics, material sciences to astrophysics and in a wide range of scales. The simplest (naive) model is to ignore conservation of mass and to view fragmentation as a multiplicative process in which the sizes of children are fractions of the size of their parents. If we assume that the succession of breaking events are independent and concentrate on a *given generation rank n*, our above result (23) applies to the distribution of fragment size X, provided we take X to zero rather than to infinity. Indeed, the factors m_1, m_2, ..., m_n are all less or equal to unity.[8] If we take, for example, $p(m) \propto \exp(-cm^a)$ for small m, we obtain $P_n(X) \propto \exp(-cnX^{a/n})$. For values of X which are order unity, large deviations theory applies when $n \to \infty$. This does not, in general, lead to a log-normal distribution, because central limit arguments are inapplicable, except in the very neighborhood of the peak of the pdf of X (see, *e.g.*, Ref. [9], sect. 8.6.5).

Next, we observe that most of the measured size distribution of fragments, *not conditioned by generation rank*, display actually power-law behavior $\propto X^{-\tau}$ with exponents τ between 1.9 and 2.6 clustering around 2.4 [19]. Several models have been proposed to rationalize these observations [13, 20] but there is no accepted theoretical description.

Here, we would like to point out a very simple and robust scenario to rationalize these observations. We again neglect the constraint that the total mass of the children is equal to that of the parent and use the simple multiplicative model. Indeed, the constraint of conservation becomes less and less important for the determination of the pdf as the generation rank increases. To illustrate what we have in mind, consider a comminution process in which, with a certain probablity less than unity, a "hammer" repetitively strikes all fragments simultaneously. Then the generation rank corresponds to the number of hammer hits. In real experiments, however, each fragment has suffered a specific number of effective hits which may

[8] When taking the logarithm, tha tail for $X \to 0$ corresponds to the régime where the sum of logarithms goes to $-\infty$. Although $X \to 0$, is not stricly speaking a "tail", we shall still keep this terminology.

vary greatly from one fragment to the other. The measurements of the size distribution should thus correspond to a superposition of pdf's of the form (23) in the tail $X \to 0$. Recent numerical simulations of lattice models with disorder [21] show indeed that, for sufficient disorder, the fragmentation can be seen as a cascade branching process.

Let us now assume that the tail of the size distribution for a fixed generation rank n is given by (23) and that the mean number $N(n)$ (per unit volume) of fragments of generation rank n grows exponentially: $N(n) \propto e^{\lambda n}$ with $\lambda > 0$. It then follows that the tail of the unconditioned size distribution is given by

$$P_{\text{size}}(X) \sim \sum_{n=0}^{\infty} [p(X^{1/n})]^n e^{\lambda n} \sim \int_0^{\infty} dn \; e^{n \ln p(X^{1/n}) + n\lambda} . \tag{27}$$

Application of Laplace's method in the variable n, treated as continuous, gives a critical (saddle) point

$$n_s = -\frac{1}{\alpha} \ln X , \tag{28}$$

where α is the solution of the transcendental equation

$$\lambda + \ln p\,(e^{-\alpha}) + \alpha e^{-\alpha} \frac{p'\,(e^{-\alpha})}{p\,(e^{-\alpha})} = 0 . \tag{29}$$

The leading-order tail behavior of the size distribution is thus given by

$$P_{\text{size}}(X) \sim X^{-\tau} , \tag{30}$$

with an exponent

$$\tau = \frac{1}{\alpha} [\ln p\,(e^{-\alpha}) + \lambda] . \tag{31}$$

This solution (30) holds for λ smaller than a threshold λ_c dependent on the specific structure of the pdf $p(x)$. For instance, consider $p(x) \propto \exp(-Cx^{\delta})$ for $x \to 0$, with $\delta > 0$. This corresponds to a pdf going to a constant as $x \to 0$, with a vanishing slope ($\delta > 1$), infinite slope ($\delta < 1$) or finite slope ($\delta = 1$). The equation (29) for α becomes $\lambda/C = (1 + \alpha\delta)e^{-\alpha\delta}$. This has a solution only for $\lambda \leq C$, as the function $(1 + x)e^{-x}$ has its maximum equal to 1 at $x = 0$. For λ approaching C from below, the exponent of the power law distribution is given by $\tau = C\delta + O(\sqrt{C - \lambda})$. At the other end $\lambda \to 0^+$, we get $\tau \to C\delta e$. In between, for $0 \leq \lambda \leq C$, the quantity $\tau/(C\delta)$ goes con-

tinuously from $e \approx 2.718$ to 1. It is interesting that τ depends on the parameters of the pdf $p(x)$ only through the product $C\delta$.

What happens for $\lambda > C$? To find out, we return to the expression (27) giving the tail of the unconditioned size distribution and find that the exponential in the integral reads $e^{n(\lambda - CX^{\delta/n})}$. In the limit of small fragments $X \to 0$, the term $X^{\delta/n}$ is dominated by the large n limit for which it is bounded by 1. Thus, $\lambda - CX^{\delta/n} \leq \lambda - C$. For $\lambda - C$, the larger n is, the larger the exponential is, while for $\lambda < C$ there is an optimal generation number n_*, for a given size X, given by (28). For $\lambda \geq C$, the critical value n_* moves to infinity. Physically, this is the signature of a shattering transition occurring at $\lambda = C$: for $\lambda > C$, the number of fragments increases so fast with the generation number n (as $e^{\lambda n} > e^{Cn}$) that the distribution of fragment sizes develops a finite measure at $X = 0$. This result is in accordance with intuition: it is when the number of new fragments generated at each hammer hit is sufficiently large that a dust phase can appear. This *shattering transition* has been obtained first in the context of mean field linear rate equations [22].

Consider another class of pdf $p(x) \propto \exp(-Cx^{-\delta})$ for $x \to 0$, with $\delta > 0$. The pdf $p(x)$ goes to zero faster than any power law as $x \to 0$ (*i.e.* has an essential singularity). The difference with the previous case is that, as the multiplicative factor $x \to 0$ occurs with very low probability in the present case, we do not expect a large number of small fragments to be generated. This should be reflected in a negative value of the exponent τ. This intuition is confirmed by an explicit calculation showing that τ becomes the opposite of the value previously calculated, *i.e.* $\tau/(C\delta)$ goes continuously from $-e \approx -2.718$ to -1 as λ goes from 0 to C.

In sum, we propose that the observed power-law distributions of fragment sizes could be the result of the natural mixing occurring in the number of generations of simple multiplicative processes exhibiting extreme deviations. This power-law structure is very robust with respect to the choice of the distribution $p(x)$ of fragmentation ratios, but the exponent τ is not universal. The proposed theory leads us to urge the making of experiments in which one can control the generation rank *of each* fragment. We then predict that the fragment distribution will not be (quasi-) universal anymore buton the contrary characterize better the specific mechanism underlying the fragmentation process.

The result (30) only holds in the "tail" of the distribution for very small fragments. In the center, the distribution is still approximately log-normal. We can thus expect a relationship between the characteristic size or peak fragment size and the tail structure of the distribution. It is in fact possible to show that the exponent τ given by (31) is a *decreasing* function of the peak fragment size: the smaller is the peak fragment size, the larger will be

the exponent (the detailed quantitative dependence is a specific function of the initial pdf). This prediction turns out to be verified by the measurements of particle size distributions in cataclastic (*i.e.* crushed and sheared rock resulting in the formation of powder) fault gouge [23]: the exponent τ of the finer fragments from three different faults (San Andreas, San Gabriel and Lopez Canyon) in Southern California was observed to be correlated with the peak fragment size, with finer gouges tending to have a larger exponent. Furthermore, the distributions were found to be a power law for the smaller fragments and log-normal by mass for sizes near and above the peak size.

5.2. *Stretched Exponential Relaxation*

We would like to suggest a possible application of the stretched exponential distribution to rationalize stretched exponential relaxations. *A priori*, we are speaking of a different kind of phenomenon: so far we were discussing distributions, while we now consider the time dependence of a macroscopic variable relaxing to equilibrium. In contrast to simple liquids where the usual Maxwell exponential relaxation occurs, "complex" fluids [24], glasses [14, 15, 25], porous media, semiconductors, *etc.*, have been found to relax with time t as $e^{-\alpha t^{\beta}}$, with $0 < \beta < 1$, a law known under the name Kohlrausch-Williams-Watts law [14, 15]. Even, the Omori $1/t$ law for aftershock relaxation after a great earthquake has recently been challenged and it has been proposed that it be replaced by a stretched exponential relaxation [26]. This ubiquitous phenomenon is still poorly understood, different competing mechanisms being proposed. An often visited model is that of relaxation by progressive trapping of excitations by random sinks [15]. Models of hierarchically constrained dynamics for glassy relaxations [25] suggest the relevance of multiplicative processes to account for the relaxation in these complex, slowly relaxing, strongly interacting materials. Our model offers a simple explanation for the difference in β measured by the same method on different materials in terms of the dependence of β on the typical number of levels of the hierarchy as we now show.

We assume that a given system can be viewed as an ensemble of states, each state relaxing exponentially with a characteristic time scale. Each state can be viewed locally as corresponding to a given configuration of atoms or molecules leading to a local energy landscape. As a consequence, the local relaxation dynamics involves a hierarchy of degrees of freedom up to a limit determined by the size of the local configuration. In phase space, the representative point has to overcome a succession of energy barriers of statistically increasing heights as time goes on; this is at the origin of the slowing

down of the relaxation dynamics. The characteristic time t_i to overcome a barrier ΔE_i is given by the Arhenius factor $t_i \sim \tau_0 e^{\Delta Ei/kT}$, where k is the Boltzmann constant, T the temperature and τ_0 a molecular time scale. For a succession of barriers increments, we get that the characteristic time is given by a multiplicative process, where each step corresponds to climbing the next level of the hierarchy. In other words, the characteristic relaxation time of a given cluster configuration is obtained by a multiplicative process truncated at some upper level. It is important to notice that our model is fundamentally different from the idea of diffusion of a representative particle in a random potential with potential barriers increasing statistically at long times, as in Sinai's anomalous diffusion [27]. We consider rather that the system can be divided into an ensemble of local configurations, each of them hierarchically ordered.

In this simple model, the times $T_n = \tau_0 (t_1/\tau_0) \ldots (t_n/\tau_0)$ are thus log-normally distributed in their center with stretched exponential tails according to our extreme deviation theory. Now, in a macroscopic measurement, one gets access to the average over the many different local modes of relaxation, each with a simple exponential relaxation: an observable $\mathcal{O}$ is thus relaxing macroscopically as $\mathcal{O} \sim \langle e^{t/Tn} \rangle$, where the average of the observable $\mathcal{O}$ is carried out over the distribution of T_n. For large t (compared to the molecular time scale), Laplace's method gives the leading-order behavior

$$\mathcal{O} \sim e^{-a t^{\beta}},$$

with $\beta = \frac{\alpha}{n+\alpha}$ for a distribution of T_n given by $e^{-an(Tn/To)^{\alpha/n}}$. In this calculation, we have assumed that all local configuration clusters are organized hierarchically according to a fixed number of n levels. We envision that this organization reflects the local atomic or molecular arrangement such that the system can be subdivised into a set of essentially mutually independent local configurations. These configurations can tentatively be identified with the locally ordered structures observed in randomly packed particles [28], macromolecules [29], glasses and spinglasses [30]. The ultrametric structure found to describe the energy landscape of the spinglass phase of mean field models also leads to a multiplicative cascade [31, 32]. Notice that if a system possesses *multiple* configuration levels n, then by the same mechanism which in fragmentation led to (27), the relaxation becomes a power law instead of a stretched exponential.

The often encountered value $\beta \approx 1/2$ corresponds, in our model, to the existence of $n \approx \alpha$ levels of the hierarchy. It is noteworthy that the factor α can be determined quantitatively from the pdf $p(t_i/\tau_0) \sim \exp[-a(t_i/\tau_0)^{\alpha}]$ of

the multiplicative factors, thus giving the potential to measure the number of levels of the hierarchy that are visited by the dynamical relaxation process. This could be checked for instance in multifragmentation in nuclear collisions, utilizing techniques sensitive to the emission order of fragments [33].

Hierarchical structures are also encountered in evolutionary processes [34], computing architectures [35] and economic structures [36] and, as a consequence, it is an interesting question whether to expect dynamical slowing down of the type described above.

5.3. *Turbulence*

In fully developed turbulence, random multiplicative models were introduced by the Russian school [37-39] and have been studied extensively since. Indeed, their fractal and multifractal properties provide a possible interpretation for the phenomenon of intermittency [40, 41] (see also ref. [9]). The pdf's of longitudinal and tranverse velocity increments clearly reveal a Gaussian-like shape at large separations and increasingly stretched exponential tail shapes at small separations, as shown in figure 1 [42-46].

Within the framework of random multiplicative models, our theory suggests a natural mechanism for the observed stretched exponential tails at small separations as resulting from extreme deviations in a multiplicative cascade. However, this mechanism cannot account for *all* properties of velocity increments. For example, random multiplicative models are not consistent with the additivity of increments over adjacent intervals. Indeed, the pdf of velocity increments δv cannot become much larger than the single-point pdf, as it would if the former were $\propto \exp(-C|\delta v|^{\beta})$ with $0 < \beta < 2$ while the latter would be close to Gaussian (see the appendix of ref. [46]). Nevertheless, stretched exponentials could be working in an intermediate asymptotic range of not too large increments, the controlling parameter of this intermediate asymptotics being the separation over which the increment is measured.

Acknowledgments

We have benefited from discussions with M. Blank and with H. Frisch. This paper is Publication no. 4711 of the Institute of Geophysics and Planetary Physics, University of California, Los Angeles.

APPENDIX

Proof of the Main Results for Extreme Deviations

Our aim is to prove (7) without necessarily assuming that the function $f(x)$, which defines the pdf of the individual variables though (4), is Taylor expandable to all orders. Specifically, we assume that f is three times continuously differentiable and satisfies the following conditions when $x \to +\infty$:

(i) $f(x) \to +\infty$ sufficiently fast to ensure the normalization (3);

(ii) $f''(x) > 0$ (convexity), where f'' is the second derivative of f;

$\widetilde{\text{(iii)}}$ $\lim \dfrac{f'''(x)}{(f''(x))^{3/2}} = 0;$

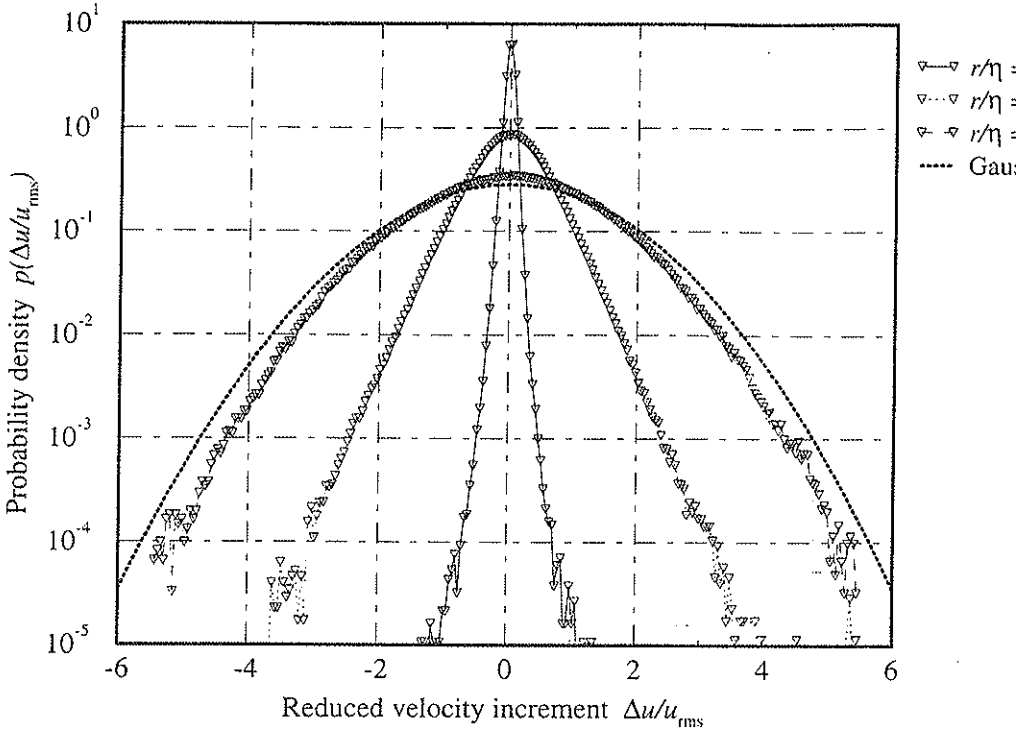


Fig. 1. pdf of transverse velocity increments reduced by the rms velocity at various separations in units of the Kolmogorov dissipation scale η. (From ref. [46]).

(iv) there exists $C_1 > 0$ such that, for $x < y$ large enough, $x^2 f''(x) / (y^2 f''(y)) < C_1$;

(v) there exist $\beta > 0$ and $C_2 > 0$ such that $x^{2-\beta} f(x) > C_2$ for large enough x.

Assumptions (i) and (ii) are just the same as made in section 2. Assumption (iii) is an instance of (iii) corresponding to the third derivative. Note that nothing is assumed about higher order derivatives. Assumptions (iv) and (v) are new and will be seen to be slight strengthenings of a corollary of (iii).[9]

We begin by proving various lemmas.

LEMMA 1. *Assumption* (iii) *implies*

$$\lim x^2 f''(x) = +\infty . \tag{A.1}$$

To prove this result, we start from (iii) which may be rewritten as

$$\lim_{x\to\infty} \frac{d}{dx} \frac{1}{\sqrt{f''(x)}} = 0 . \tag{A.2}$$

It follows that for any $\epsilon > 0$ there exists $X(\epsilon)$ such that, for $x > X(\epsilon)$, the argument of the limit in (A.2) is less than ϵ in absolute value. We take $X(\epsilon) < y < x$ and apply the mean value theorem to get

$$\left| (f''(x))^{-1/2} - (f''(y))^{-1/2} \right| = (x-y) \left| \frac{d}{d\xi} (f''(\xi))^{-1/2} \right| , \tag{A.3}$$

with $y < \xi < x$. The rhs of (A.3) is less then $(x-y)\epsilon$. Dividing by x and letting $x \to \infty$, we obtain

$$\limsup_{x\to\infty} \frac{d}{x\sqrt{f''(x)}} \leq \epsilon . \tag{A.4}$$

Letting $\epsilon \to 0$, we obtain (A.1). QED.

[9] There are weaker formulations of (iv) and (v) for which our results hold, which we do not make explicit here, as they are quite involved and do not bring any additional insight.

LEMMA 2. *Under assumptions* (ii), (iii)~ *and* (iv),

$$|x-y| = C(f''(x))^{-1/2} \tag{A.5}$$

implies, that

$$\frac{f''(y)}{f''(x)} \to 1, \quad \textit{for } x \to +\infty \textit{ and fixed } C. \tag{A.6}$$

For the proof, let us first assume that $y < x$. By the mean value theorem, we have

$$f''(x) - f''(y) = (x-y)f'''(\xi), \quad \text{with } y < \xi < x\,. \tag{A.7}$$

It follows from Lemma 1 and (A.5), that $x/y \to 1$ and thus $\xi/x \to 1$ as $x \to +\infty$. Dividing (A.7) by $f''(x)$ and using (A.5), we obtain

$$\frac{f''(x)-f''(y)}{f''(x)} = C\,\frac{f'''(\xi)}{(f''(\xi))^{3/2}}\left(\frac{f''(\xi)}{f''(x)}\right)^{3/2}. \tag{A.8}$$

By (iv), the rightmost factor on the rhs is less than $C_1^{3/2}(x/\xi)^3$, which remains bounded as $x \to +\infty$, while, by (iii)~, the leftmost factor on the rhs tends to zero. Hence, the rhs tends to zero. This implies (A.6).

For the case $x < y$, (A.7) holds similarly with $x < \xi < y$. We then multiply (A.7) by $(f''(x))^{1/2}/(f''(y))^{3/2}$. The rhs tends again to zero. If follows that

$$\frac{f''(x)-f''(y)}{f''(y)}\left(\frac{f''(x)}{f''(y)}\right)^{1/2} \to 0\,, \tag{A.9}$$

which implies again (A.6). QED.

LEMMA 3. *Let* h_i, $i = 1, \ldots, n$ *be real variables, not all vanishing, such that* $\sum_{i=1}^{n} h_i = 0$. *The subset of* $p \le n-1$ *indices* i_j *such that* $h_{i_j} \ge 0$ *satisfes*

$$\sum_{j=1}^{p} h_{i_j}^2 \ge \frac{1}{n}\sum_{i=1}^{n} h_i^2\,. \tag{A.10}$$

Let $i_{p+1}, \ldots, i_n$ denote the subset of indices such $h_{i_j} < 0$. We set $h'_{i_j} = -h_{i_j} > 0$, so that

$$\sum_{j=1}^{p} h_{i_j} = \sum_{j=p+1}^{n} h'_{i_j} . \tag{A.11}$$

We have

$$\begin{aligned} \sum_{j=1}^{n} h_i^2 &= \sum_{j=1}^{p} h_{i_j}^2 + \sum_{j=p+1}^{n} h_{i_j}'^2 \\ &\leq \sum_{j=1}^{p} h_{i_j}^2 + \left(\sum_{j=p+1}^{n} h'_{i_j} \right)^2 \\ &= \sum_{j=1}^{p} h_{i_j}^2 + \left(\sum_{j=1}^{p} h_{i_j} \right)^2 \\ &\leq \sum_{j=1}^{p} h_{i_j}^2 + p \sum_{j=1}^{p} h_{i_j}^2 \\ &\leq n \sum_{j=1}^{p} h_{i_j}^2 . \end{aligned} \tag{A.12}$$

In deriving (A.12), we have used $p \leq n - 1$ and the following inequality for a set of p nonnegative variables $y_1, \ldots, y_p$

$$(y_1 + \cdots + y_p)^2 \leq p(y_1^2 + \cdots + y_p^2) . \tag{A.13}$$

Lemma 3 follows from (A.12). QED.

We now turn to the derivation of the main result (7), rewritten here in a slightly different form as:

$$\lim_{x \to +\infty} \frac{P_n(x)}{P_n^{\text{as.}}(x)} = 1 , \tag{A.14}$$

where

$$P_n^{\text{as.}}(x) = \mathrm{e}^{-nf(x/n)} \frac{1}{\sqrt{n}} \left(\frac{2\pi}{f''(x/n)} \right)^{\frac{n-1}{2}} . \tag{A.15}$$

We start from the representation (12) of the pdf the sum of n iid variables as an $(n-1)$-fold integral. The function g_n, given by (11) can be rewritten as

$$g_n = nf(x/n) + \sum_{i=1}^{n} \int_{\frac{x}{n}}^{\frac{x}{n}+h_i} dz \int_{\frac{x}{n}}^{z} f''(y)dy \,. \tag{A.16}$$

Observe that, by (10), we have $\sum_{i=1}^{n} h_i = 0$, so that, ignoring contributions of zero measure, at least one of the h_i's must be positive. Furthemore, all the terms involving double integrals are positive.

The proof goes now as follows. By Lemma 2, the second derivative $f''(y)$ can be replaced by the second derivative at the minimizing point x/n as long as all the h_i's are not too large, that is are in the set $\mathcal{A}_H$ defined by

$$|h_i| \leq H = C(f''(x/n))^{-1/2}, \text{ for all } i \,. \tag{A.17}$$

By (A.1), (A.17) expresses that all the individual random terms in the sum stay within a distance of x/n which is small compared to x, that is, what we have called the *democratic localization* property. The substitution of $f''(x/n)$ for $f''(y)$ amounts to using the second-order truncation of the Taylor series (13) for g_n, which leads to $P_n^{\text{as.}}(x)$. It follows from Lemma 2 that the error committed in this substitution is small for large x.

Since $f''(x) > 0$, the contribution of the complementary set $\overline{\mathcal{A}_H}$ to the pdf $P_n(x)$, denoted $P_n^{(>H)}(x)$, is estimated from above by estimating g_n from below, keeping only the contributions from the subset i_j $(j = 1, \ldots, p \leq n-1)$ of indices such that $h_{i_j} \geq 0$. We thus obtain

$$g_n \geq nf(x/n) + \sum_{j=1}^{p} \int_{\frac{x}{n}}^{\frac{x}{n}+h_{i_j}} dz \int_{\frac{x}{n}}^{z} f''(y)\, dy \,. \tag{A.18}$$

By (iv), for $x/n \leq y \leq x/n + h_{i_j}$, we have

$$y^2 f''(y) \geq C_1^{-1} (x/n)^2 f''(x/n) \,. \tag{A.19}$$

Using (A.19) in (A.18), we obtain

$$g_n \geq nf(x/n) + C_1^{-1} \left(\frac{x}{n}\right)^2 f''(x/n) \sum_{j=1}^{p} q(nh_{i_j}/x) \,, \tag{A.20}$$

where

$$q(\alpha) \equiv \alpha - \ln(1+\alpha)\,. \tag{A.21}$$

Note that $q(\alpha) = \alpha^2/2 + O(\alpha^3)$ for small α and $q(\alpha) < \alpha$ for large α. Assumption (v) is used to show that, for large x, the overwhelming contribution to $P_n^{(>H)}(x)$ comes from h_{ij}'s such that nh_{ij}/x is small compared to unity. Using (A.20) and Lemma 3, we obtain the following estimate

$$P_n^{(>H)}(x) \leq e^{-nf(x/n)} \underbrace{\int \cdots \int}_{n-1} e^{-\frac{C_1^{-1}}{2n} f''(x/n) \sum_{i=1}^{n} h_i^2}\, dh_1 \ldots dh_{n-1}\,, \tag{A.22}$$

where the domain of integration is over $\overline{\mathcal{A}_H}$, so that at least one of the $|h_i| \geq H = C(f''(x/n))^{-\frac{1}{2}}$. As a consequence, it is easily checked that the bounding integral is less than $P_n^{\text{as.}}(x)$ multiplied by a factor $O(e^{-C^2/n})$, which tends to zero very quickly for large C. This proves (A.14) and the democratic localization property.

REFERENCES *

[1] Gnedenko B.V. and Kolmogorov A.N., Limit distributions for sums of independent random variables (Addison Wesley, Reading MA, 1954).

[2] Feller W., An introduction to probability theory and its applications, vol. II (John Wiley and sons, New York, 1971).

[3] Cramér H., *Actulalités Sci. Indust.* **736** (1938) 5-23.

[4] Varadhan S.R.S., Large Deviations and Applications (SIAM, Philadelphia, 1984).

[5] Ellis R.S., Entropy, Large Deviations and Statistical Mechanics (Springer, Berlin, 1985).

[6] Lanford O.E., Entropy and equilibrium states in classical mechanics, in "Statistical Mechanics and Mathematical Problems", A. Lenard, Ed. (Springer, Berlin) *Lect. Notes Phys.* **20** (1973) 1-113.

[7] Jensen J.L., Saddlepoint Approximations (Oxford Science Publications, Clarendon Press, Oxford, 1995).

[8] Bender C. and Orszag S.A., Advanced Mathematical Methods for Scientists and Engineers (McGraw-Hill, New York, 1978).

[9] Frisch U., Turbulence: the Legacy of A.N. Kolmogorov (Cambridge University Press, 1995).

[10] Broniatowski M. and Fuchs A., *Adv. Math.* **116** (1995) 12-33.

[11] Borovkov A.A. and Mogulskii A.A., *Sib. Adv. Math.* **2** (1992) 52-120.

[12] Esscher F., *Skand. Akctuarietidskrift* (1932) p. 175.

[13] Redner S., Fragmentation, in *Statistical models for the fracture of disordered media*, H.J. Herrmann and S. Roux, Eds. (Elsevier Science Publishers, 1990); Cheng Z. and Redner S., *Phys. Rev. Lett.* **60** (1988) 2450-2453; Ouillon G., Sornette D., Genter A. and Castaing C., *J. Phys. I France* **6** (1996) 1127-1139.

[14] Gielis G. and Maes C., *Europhys. Lett.* **31** (1995) 1-5; Chung S.H. and Stevens J.R., *Am. J. Phys.* **59** (1991) 1024-1030; Alvarez F., Alegria A. and Colmenero J., *Phys. Rev. B* **44** (1991) 7306-7312.

[15] Phillips J.C., *Rep. Prog. Phys.* **59** (1996) 1133-1208.

[16] Ghashghaie S., Breymann W., Peinke J., Talkner P. and Dodge Y., *Nature* **381** (1996) 767; Ghashghaie S., Breymann W., Peinke J. and Talkner P., Turbulence and financial markets, in "Proceedings European Turbulence Conference VI", *Advances in Turbulence VI*, S. Gavrilakis, L. Machiels and P.A. Monkewitz, Eds. (Kluwer, 1996) pp. 167-170.

[17] Arnéodo A., Bouchaud J.-P., Cont R., Muzy J.-F., Potter M. and Sornette D., Comment on "Turbulent cascades in foreign exchange markets" (cond-mat/9607120) (reply to Ghashghaie *et al.*, 1996); Mantegna R.N. and Stanley H.E., Stock market dynamics and turbulence: parallels in quantitative measures of fluctuation phenomena, preprint (1995); Turbulence and financial markets, *Nature* (Scientific Correspondence) **383** (N6601) (1996) 587-588; Arnéodo A., Muzy J.-F. and Sornette D., Causal cascade in the stock market from the "infrared" to the "ultraviolet", *Nature*, submitted.

---

* Copyeditor's note. As this paper comes directly from a previously published article, it has been decided to leave the technical apparatus in its original format.

[18] Mantegna R. and Stanley H.E., *Nature* **376** (N6535) (1995) 46-49.

[19] Turcotte D.L., *J. Geophys. Res.* **91** (B2) (1986) 1921-1926.

[20] Marsili M. and Zhang Y.C., *Phys. Rev. Lett.* **77** (1996) 3577-3580.

[21] Astrom J. and Timonen J., *Phys. Rev. Lett.* **78** (1997) 3677-3680.

[22] Maslov D.L., *Phys. Rev. Lett.* **71** (1993) 1268-1271; Boyer D., Tarjus G. and Viot P., *Phys. Rev. E* **51** (1995) 1043-1046.

[23] An L.-J. and Sammis C.G., *Pageoph.* **143** (1994) 203-227.

[24] Klinger M.I., *Phys. Rep.* **165** (1988) 275-397.

[25] Palmer R.G., Stein D.L., Abrahams E. and Anderson P.W., *Phys. Rev. Lett.* **53** (1984) 958.

[26] Kisslinger C., *J. Geophys. Res.* **98** (1993) 1913-1921.

[27] Bouchaud J.-P. and Georges A., *Phys. Rep.* **195** (1990) 127-293.

[28] Mosseri R. and Sadoc J.F., *J. Phys. Lett. France* **45** (1984) L-827.

[29] Levitt M., *Ann. Rev. Biophys. Bioeng.* **11** (1982) 251.

[30] Mézard M., Parisi G. and Virasoro M.A., Spinglass Theory and Beyond, *World Scientist Lecture Notes in Physics*, Vol. 9 (1987).

[31] Bouchaud J.-P. and Dean D.S., *J. Phys. I France* **5** (1995) 265-286.

[32] Saleur H. and Sornette D., *J. Phys. I France* **6** (1996) 327-355.

[33] Cornell E.W. *et al.*, *Phys. Rev. Lett.* **77** (1996) 4508-4511.

[34] O'Neill R.V. *et al.*, A Hierarchical Concept of Ecosystems (Princeton University Press, Princeton, N.J., 1986).

[35] Huberman B.A. and Kerszberg M., *J. Phys. A* **18** (1985) L331.

[36] Tostesen E., Dynamics of hierarchically clustered cooperative agents, Cand. Scient. Thesis, University of Copenhagen (1995).

[37] Kolmogorov A.N., *J. Fluid Mech.* **13** (1962) 82-85.

[38] Yaglom A.M., *Dokl. Akad. Nauk SSSR* **166** (1966) 49-52.

[39] Novikov E.A. and Stewart R.W., *Izv., Akad. Nauk SSSR, Ser. Geoffiz.* (1964) pp. 408-413.

[40] Mandelbrot B., *J. Fluid Mech.* **62** (1974) 331-358.

[41] Parisi G. and Frisch U., On the singularity structure of fully developed turbulence, in "Turbulence and Predictability in Geophysical Fluid Dynamics", Proceed. Intern. School of Physics 'E. Fermi', 1983, Varenna, Italy, M. Ghil, R. Benzi and G. Parisi, Eds. (North-Holland, Amsterdam, 1985) pp. 84-87.

[42] Vincent A. and Meneguzzi M., *J. Fluid Mech.* **225** (1991) 1-25.

[43] Zocchi G., Tabeling P., Maurer J. and Willaime H., *Phys. Rev. E* **50** (1994) 3693-3700.

[44] Kahalerras H., Malecot Y. and Gagne Y., Transverse structure functions in three-dimensional turbulence, in "Advances in Turbulence" VI, S. Gavrilakis, L. Machiels and P.A. Monkewitz, Eds. (Kluwer, 1996) pp. 235-238.

[45] Herweijer J.A. and Van der Water W., Transverse structure functions of turbulence, in "Advances in Turbulence" V, R. Benzi, Ed. (Kluwer, 1995) pp. 210-216.

[46] Noullez A., Wallace G., Lempert W., Miles R.B. and Frisch U., *J. Fluid Mech.* **339** (1997) 287-307.

CONSILIENCE FROM RIVER NETWORKS

ANDREA RINALDO

INTRODUCTION

The complexity of interactions between economics, society and the environment has motivated, in recent years, the need for defining and pursuing sustainable development, i.e. a blend of economic growth that maintains adequate standards for the quality of life and of the environment. This requires the definition of a background or reference state of nature, meant as the reference ecosystem, a desirable state in which the natural evolutionary dynamics would settle in some stable manner. But what if such a state of the environment does not exist? How do we cope with ecosystems, or inanimate open, dissipative systems with many degrees of freedom like river networks or lagoons, that may have no preferential state in their evolutionary dynamics? The above questions are, in my view, the paradigm of some key choices that humankind will have to face in the near future – should we pursue social and economic prosperity, the strict physical preservation, maybe under glass bells, of the current configuration of a complex system to preserve some its current features, or should we allow unleashed natural evolution and risk the loss of the features we treasure? In some yet wider sense, does the notion of natural equilibrium, in the static maintenance sense that it somewhat suggests, make any sense in natural evolutionary phenomena like those at work in the dynamics of large, open, dissipative systems so ubiquitous in nature?

The above questions lie at the heart of the theme chosen for this session, centered on criticality and self-organization. One clear, practical example of global significance is the survival of the city of Venice (Musu, 1998), an issue of paramount importance not only because of the intrinsic importance of the historic, artistic and architectural value of the city, but also because it is a paradigm of the complexity of the interactions among economics, society and the environment.

In this paper I analyse with some hindsight how nature works to produce river networks (e.g. Rodriguez-Iturbe and Rinaldo, 1997). Three factors did come together to effect our learning of her lesson. The first is the appearance on the scientific scene of Mandelbrot's (1983) ideas on fractal geometry. The recognition and the implications of the fractal geometry of nature have radically changed the way we perceive and measure natural phenomena. The second factor is the emergence of a coherent theoretical framework for the dynamic origin of fractals, chiefly theories of critical self-organization (e.g. Bak, 1993). Theory thus provided a broad foundation to the linkages of scale-invariant forms, so common both in the living (e.g. McMahon and Bonner, 1983) and the inanimate world (Mandelbrot, 1983; Niemeyer *et al.*, 1983; Sander, 1987) with general features of the embedded evolutionary dynamics. The third factor is the accessibility of large data sets, objectively collected, of acceptable precision and spanning the natural phenomena over a wide range of scales (Rodriguez-Iturbe and Rinaldo, 1997; Rinaldo *et al.*, 1998). Such factors, owing to digital mapping techniques, provided hydrologists and geomorphologists with a unique opportunity for the analysis and the testing. As we now fully realize, many geomorphological relationships empirically known – some dating from the last century – carry the signatures of fractal growth and of critical self-organization (Mandelbrot, 1983) and, interestingly, of climatic changes (Rinaldo *et al.*, 1995).

River networks are akin to an important role in the transfer of ideas and results among scientific disciplines, which I like to relate to the very concept of consilience (Wilson, 1998). In the master's words, a balanced perspective cannot be acquired by studying disciplines in pieces, but through pursuit of the consilience, or the linking of facts or fact-based theory across disciplines to create a common ground of explanation. Thus if nature speaks consistently the same language and shows the same basic mechanisms, then the ubiquitous occurrence of networks in nature should be read with the same tools, and looking for a unifying principle (e.g. Kauffman, 1993). This could indeed be, as I will try to show, the Darwinian idea of the survival of the fittest. Fitness describes the probability of a global state, related to its energy features and optimal networks conform beautifully to the features we observe in nature in the compelling case of rivers (e.g. Rodriguez-Iturbe and Rinaldo, 1997). Possible implications on the dynamic origin of scale invariance from the general properties of transportation networks (e.g. Stevens, 1974; Ball, 1999), including circulatory or respiratory netwoks in living organisms, will be briefly described. It is thus suggested that river networks might help bridge the gap between physics and biology, and address size-rate relationships that characterize biological processes from cellular metabolism to population dynamics.

On River Networks

Lasting accomplishments in geomorphology (Leopold and Langbein, 1962; Shreve, 1966, 1967) have dealt with the study of random-walk and topologically random channel networks. Through the random perspective, which has had a profound influence on the interpretation of natural landforms, nature's resiliency in producing recurrent networks and landforms was interpreted to be the consequence of chance. In fact, central to models of topologically random networks is the assumption of equal likelihood of any tree-like configuration. However, there exists a general framework of analysis which argues that all possible network configurations draining a fixed area are not necessarily equally likely. Rather, a probability $P(s)$ is assigned to a particular spanning tree configuration, say s, which can be generally assumed to obey a Boltzmann distribution: $P(s) \propto e^{-H(s)/\mathcal{T}}$ where $\mathcal{T}$ is a parameter and $H(s)$ is a global property of the network configuration s related to energetic characters, i.e. its Hamiltonian. One extreme case is the random topology model where all trees are equally likely, i.e. the limit case for $\mathcal{T} \to \infty$. The other extreme case is $\mathcal{T} \to 0$ and this corresponds to network configurations that tend to minimize their total energy dissipation to improve their likelihood. Networks obtained in this manner are termed Optimal Channel Networks (OCNs) (Rodriguez-Iturbe and Rinaldo, 1997).

It should be noted that herein I address channel networks, that is, attention is focused on planar landforms generated by fluvial erosion processes acting on evolving landscapes. Not that these are unrelated to the three-dimensional structure of the landscape – drainage directions are identified through local topographic gradients. One cannot, however, uniquely (or always) relate a planar aggregation structure to a suitable altitude field. Nevertheless attention may only be confined to planar landforms and statistics without loss of generality (Rodriguez-Iturbe and Rinaldo, 1997).

One crucial feature of the organization of scale-invariant structures is the power-law structure of the probability distributions characterizing their geometrical properties (Mandelbrot, 1967, 1983). This behaviour, characterized by events and forms of all sizes, is consistent with the fact that many complex systems in nature evolve in an intermittent, burst-like way rather than in a smooth, gradual manner. For spatial structures like river networks, mountain ranges or coastlines, this implies a power-law probability distribution of geometric quantities like total contributing area at a site, of stream lengths among others. These properties are ubiquitous in nature. The distribution of earthquake magnitudes obeys Gutemberg-Richter's law (1956) which is a power-law of energy release. Fluctuations in economics also follow power-law distributions with long tails describing intermittent

Fig. 1. One example of natural river network extracted from high accuracy digital terrain maps. The general idea is that without a scale bar it is impossible to devise even the approximate scale of a planar river form or of a topographic map (Mandelbrot, 1983), an ingredient essential to any form of scale-invariance. Although this is true for a wide range of scales, there exist a lower and an upper limit to landscape dissection into distinct valleys. The lower cutoff is generated by a threshold of channelization that sets a finite scale to the landscape. The upper cutoff is usually of a geological nature. Nevertheless it is remarkable that scale invariance holds from the order of 10 metres to the tens of thousands of kilometres (e.g. Rodriguez-Iturbe and Rinaldo, 1997).

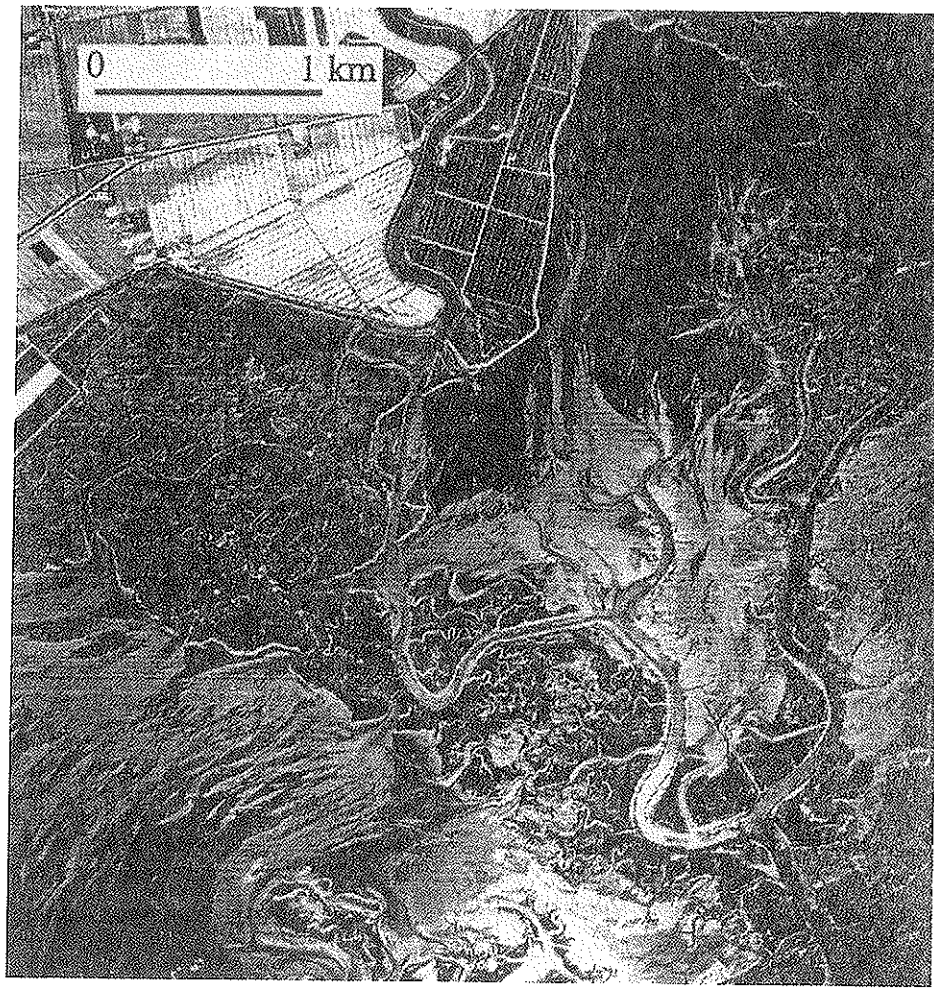


Fig. 2. A particular of tidal networks (river-like channels generated in hydrodynamic basins where tidal forcing generates pulsatile flow that cyclically inverts direction and peaks approximately at the same rates in ebb and flood) in the northern lagoon of Venice (Italy). The beauty of landforms is awesome, yet they are bound to disappear in a few hundred years in connection with the natural evolution of lagoonal environments in sub-siding scenarios. What strikes is the diversity of network forms that develop within a very short range. One notes features quite different from those typical of fluvial networks, typically related to the strong longitudinal gradients of flowrates induced by the hydrodynamic of lateral storage regions (*barene, velme*) and their incisions. Nonetheless a strive towards some form of scale invariance is shown regardless of a radically different dynamic context. The similarities and the differences of tidal and river networks may indeed shed light on the characters of pulsatile and nonpulsatile networks of the living world.

large events, as first elucidated by Mandelbrot's (1963) famous example of the variation of cotton prices. Punctuations dominate biological evolution (Gould and Eldredge, 1993) where many species become extinct and new species appear interrupting periods of stasis. Coastlines reveal similar shapes when observed at different scales and through different resolutions, as – with some distinction – mountain profiles, and for a wide range of scales without a scale bar one cannot distinguish a large river network from a small one. The embedded scale invariance apparent in much of the geometry of nature can be described through linked scaling exponents (describing, in the case at hand, aggregation, lengths and elongation) which, following Maritan *et al.* (1996a) and Rigon *et al.* (1996) we will take as distinctive tools for the comparison of different channel networks. In fact, a far from superficial question that one addresses when comparing channel networks is what uniquely characterizes them.

In addressing the common properties of a tree, many of the properties of interest are surrogated by the organization of total contributing areas. In a river network the total contributing area of a site i, i.e. A_i, is identified through drainage directions and is measured by the number of sites upstream of the site connected by the network.[1] Thus at any site of a network one has a value of A. In real networks it turned out to be a random variable described by a finite-size scaling probability distribution with scaling exponent β.[2] The coefficient β characterizes the entire aggregation pattern. The observational result of Rodriguez-Iturbe *et al.* (1992a) is

$$\beta = 0.43 \pm 0.02 \tag{2}$$

which is an important character relatively independent of size, vegetation, geology, climate or orientation of the basin. Notice that the value of the slope β is unaffected by the size of the support threshold used in the iden-

---

[1] If $nn(i)$ are the nearest-neighbours to i, the equation for total contributing area is $A_i = \sum_{j \in nn(i)} A_j W_{ij} + 1$ where W_{ij} is a matrix which has value 1 if i collects flow from its neighbour j through a drainage direction (i.e. if $j \rightarrow i$), and 0 otherwise. The unit area added in eq. (1) is the contribution of the i-th pixel.

[2] The distribution is:

$$p(a) \propto a^{-(1+\beta)} \mathcal{F}(a/L_{\parallel}^{1+H}) \tag{1}$$

which has probability of exceedence of the form $P[A \geq a] \propto a^{-\beta} \mathcal{F}(a/L_{\parallel}^{1+H})$ (Rodriguez-Iturbe *et al.*, 1992a; Maritan *et al.*, 1996a) where a is the arbitrary value of the random variable A, $L_{\parallel}$ is a linear size characteristic of the network. Technically $L_{\parallel}$ is defined as the longest distance from a site on the boundary to the outlet, $\mathcal{F}$ is a scaling function whose detailed form is irrelevant but whose limits allow important comparisons, and H is a scaling exponent characterizing the elongation of the total area occupied by the network.

tification of the network, of course within limits as care should be exerted in actually dealing with channelized pixels rather than with the contributing hillslopes. The coefficient H in real basins is in the range $0.75 - 0.80$ (Rodriguez-Iturbe and Rinaldo, 1997).[3]

We also measure important properties of the length of the channels. It is appealing to define stream length properties without linking the analysis to any particular ordering scheme. One way to do this is to analyze the random variable defined as the longest distance measured through the network from a randomly chosen point to the boundary of the basin, i.e. the mainstream length L extended to the subbasins rooted in all sites of the network. Technically one defines the mainstream pattern upstream of any junction following the site having maximum area (in case of equal contributions one chooses at random) until a source is reached. The finite-size distribution of mainstream lengths L at any point in the basin for a given area A has scaling exponent ξ (Rigon *et al.*, 1996) and obeys the same requirements of that of areas.[4] The ξ scaling exponent is in the observational range is $0.75 - 0.90$ (Maritan *et al.*, 1996a; Rigon *et al.*, 1996). Notice that consistency requires that $\xi = \beta/h$.[5] Experimental evidence strongly supports the above assumptions (Rodriguez-Iturbe and Rinaldo, 1997).

In conclusion, a set of independently-measured and theoretically linked scaling exponents distinctively characterize a river network. They define the topological structure through the properties of its aggregation, the length of the individual components not seen by topology alone, the degree of irregularity of the mainstream and the general elongation of the drainage area. Thus we will assume that a synthetic shape matches natural features when all scaling exponents fall in the following ranges:

---

[3] From the previous finite-size argument Maritan *et al.* (1996a) also obtained the exact result $< A^n > \propto L_\parallel^{(n-2-\beta)(1+H)}$ which allows the introduction of the relationship

$$\frac{< A^n >}{< A^{n-1} >} \propto L_\parallel^{1+H} \tag{3}$$

which allows a direct calculation of H from data.

[4] The distribution is shown to be in the form $P[L \geq l] = l^{-\xi}\, \mathcal{G}(\frac{l}{A^h})$, where P is the probability distribution exceeding the arbitrary length l given a drainage area A, $\mathcal{G}$ is a scaling function whose specialization is immaterial.

[5] Notice also, in analogy with the previous result, that one has $< L^n > \propto A^{h(n-\xi-2)}$, i.e. the n-th moment of the distribution of mainstream lengths for a given area A scales with area A (Rigon *et al.*, 1996), and

$$\frac{< L^n >}{< L^{n-1} >} \propto A^h \tag{4}$$

(see, for details, Rodriguez-Iturbe and Rinaldo (1997, pp. 182-195)).

$$\beta = 0.43 \pm 0.02 \qquad (5)$$
$$H = 0.75 - 0.80$$
$$h = 0.57 - 0.60$$
$$\xi = = 0.8 - 0.9$$

Chance-dominated, random walk type of models, once thought to capture the essence of Nature's mechanisms, do not reproduce natural features. A complete and revealing comparison of the scaling coefficients for Leopold/Eden (Leopold and Langbein, 1963; Leopold *et al.*, 1964; Eden, 1961; Howard, 1971) and Scheidegger (1963) networks with observational data is shown in table 1. Shreve's (1966, 1967) topological random networks also do not reproduce well the scaling exponents, although they reproduce less stringent statistics.

Optimal channel network (OCN) configurations are obtained minimizing the total rate of energy expenditure, say $H(s)$, in the system as a whole and in its parts. The idea that patterns in nature may be obtained by optimality principles of an energetic character is not new (e.g. Stevens, 1974). Along these lines Rodriguez-Iturbe *et al.* (1992b) have suggested new local and global optimality principles linking energy dissipation and runoff production with the 3D structure of river basins. The derivation of the basic functional $H(s)$ to be minimized may follow different paths. The general expression of total energy dissipation, which we term $H(s)$ to suggest its Hamiltonian meanung, is

$$H(s) = \sum_i A_i^{\gamma} \qquad (6)$$

Table 1. *Summary table of scaling coefficients for change-dominated networks* (Leopold/Eden and Scheidegger) *comparted with observational data* (after Rodriguez-Iturbe and Rinaldo, 1997).

Exponent	Scheidegger trees	Leopold/Eden Networks	Real Basins
β	1/3	0.52 ± 0.02	0.43 ± 0.02
H	1/2	0.50	$0.75 - 0.80$
h	2/3	0.57	$0.57 - 0.60$
ξ	1.0	0.91 ± 0.05	$0.8 - 0.9$

where $\gamma = 1 - \theta \approx 0.5$ (see, for various related issues, Rodriguez-Iturbe and Rinaldo, 1997). Thus the general operational problem of OCNs is to find the spanning network configuration s that minimizes $H(s)$ in eq. (5). Once γ is fixed, thus defining the dominant transport process, no parameter is involved in the search thus complying with the requisites of robustness outlined in the introduction.

The optimal channel network (OCN) configuration, say s, is thus derived by the global principle of minimum energy expenditure in the network as a whole, yielding the condition on the functional determining the total energy expenditure of the system, i.e. its Hamiltonian $H(s)$:

$$H(s) = \sum_i A_i^\gamma = \text{minimum} \tag{7}$$

where i spans all the links developing in the s configuration of the network. Note that when convenient we might interchange the notations $H(s) = H_\gamma(s)$. Notice that once γ is assumed defining the dominant transport process, no parameter is involved in the search for OCNs defined by eq. (6).

The basic operational problem to obtain OCNs for a given area is to find the connected path draining it that minimizes $H(s)$ without postulating predefined features, like e.g. the number of sources or the link lengths like in topologically random networks. The reader is referred to Rodriguez-Iturbe and Rinaldo (1997) for a discussion of various related issues. The search may or may not choose a path that gets rid of the imprinting of initial conditions (Metropolis *et al.*, 1957; Kirkpatrick *et al.*, 1963). The former procedure aims at the global minimum, which we term the ground state. Any procedure following the second path (typically a greedy search procedure in which one accepts random change sonly whether they lower the Hamiltonian) is defined to lead to an OCN. At this point we recall the important result of Maritan *et al.* (1996b) who have found exactly the features of the configuration of the ground state Such features are:

$$\begin{aligned} \beta &= 0.5 \\ H &= 1 \\ h &= 0.50 \\ \xi &= 1 \end{aligned} \tag{8}$$

with which the suitable numerical experiments comply, and differ noticeably from the values observed in nature. Table 2 shows the much better performance of local minima (OCN) where an imperfect procedure of minimum search gets trapped.

Table 2. *Summary table of scaling coefficients for the global optimum* (after Maritan *et al.*, 1996b) *OCNs* (adapted from Rodriguez-Iturbe and Rinaldo, 1997) *and observational data* (see Table 1).

Exponent	Ground state	Optimal Channel Networks	Real Basins
β	0.5	0.43 ± 0.02	0.43 ± 0.02
H	1.00	0.75 ± 0.01	$0.75 - 0.80$
h	0.50	0.57 ± 0.02	$0.57 - 0.60$
ξ	1.0	0.8 ± 0.05	$0.8 - 0.9$

What are the implications of the worse energetic performance and yet the better representation of natural networks performed by OCNs? An explanation might come from the role of constraints in the evolution of a channel network. Channel networks cannot change freely, i.e. regardless of initial conditions, because these conditions have an imprinting on their evolution. One valid question is whether the optimization that nature seems to perform in the organization of the parts and the whole of the river basin can be farsighted, i.e. capable of evolving in a manner that completely disregards initial conditions and allows for major migration of divides in the search for a more stable configuration evolving through transient unfavourable conditions. The experiment described before seem to indicate that the type of optimization that nature performs is myopic, that is, willing to accept changes only if their impact is favourable right after their occurrence (in the immediate) rather than in the long run. In other words, it seems that the natural process of network evolution is unlikely to allow for dramatic changes of the initial configuration which in the short run will decrease the fitness of the configuration but which will provide paths for a later improvement leading to an overall better state.

Notice that the process of network evolution in nature is likely to move from some sort of an initial condition generated by the rapid formation of a rudimentary network created by erosion processes characterized by much faster time scales than those of mature network establishment. Hence the relatively slow process of long-term erosion balancing diffusive and concentrative transport of sediment is likely to move from the almost instantaneous creation of an early drainage structure which leaves geomorphic signatures (e.g. Rinaldo *et al.*, 1995). Geologic controls may also play the role of the (finite) size of the computational lattices.

The claim that river networks are not free to explore extended regions of their fitness landscapes, suggests that nature might not search for global minima when striving for optimality. To sustain this view, we observe that feasible optima such as dynamically accessible structures have also been found – motivated by channel networks – for interfaces of random ferromagnets and for stable dendritic structures developing in potential electric fields (see, for a review, Rodriguez-Iturbe and Rinaldo, 1997). Many more examples are conceivable, and many cases where the suboptimal states differ from the ground state are known. The peculiarity of channel networks is that they show a richness of diverse yet statistically consistent scale-invariant structure for all suboptimal states, which complete Shreve's view in that nature's resiliency is produced by a large number of possible similar states, though the mechanism that decides the dynamic accessibility is not driven by chance alone. Thus we deem reasonable that complex proteins fold into dynamically accessible states just as river networks or interfaces of ferromagnets adjust to feasible shapes, and the emergence of scale-invariant structures is tied, in our interpretation, to the establishment of global interactions constraining the system in a dynarnic context searching for optimality.

Acknowledgements

It is a pleasure to acknowledge my intellectual partnerships and the many contributions of Jayanth Banavar, Penny Chisholm, Marco Marani, Amos Maritan, Riccardo Rigon and Ignacio Rodriguez-Iturbe, which have been central to the formation of the ideas contained in this paper.

REFERENCES

Ball, P. (1999): *The Self-Made Tapestry: Pattern Formation in Nature* (Oxford, Oxford University Press).

Bak, P. (1996): *How Nature Works: the Science of Self-Organized Criticality* (New York, Copernicus-Springer).

Dietrich, W.E., Wilson, C.J., Montgomery, D.R., McKean, J., and Bauer, R. (1992): 'Erosion Thresholds and Land Surface Morphology', *J. Geology, 20*, pp. 675-679.

Eden, M. (1961): 'A Two-dimensional Growth Process, Fourth Berkeley Symposium on Mathematical Statistics and Probability', in R. Neyman (ed.), *Biology and Problems of Health* (Berkeley, Univ. of California Press), pp. 223-239.

Gould, S.J., and Eldredge, R. (1993): 'Punctuated Equilibrium Comes of Age', *Nature, 366*, pp. 223-226.

Hack, J.T. (1957): 'Studies of Longitudinal Profiles in Virginia and Maryland', *U.S. Geological Survey Paper 294-B*, Washington.

Howard, A.D. (1971): 'Simulation Model of Stream Capture', *Geol. Soc. Am. Bull., 82*, pp. 1355-1363.

Kauffman, S. (1993): *The Origins of Order* (New York, Oxford University Press).

Kirkpatrick, S., Gelatt, G.D., and Vecchi, M.P. (1983): 'Optimization by Simulated Annealing', *Science, 220*, pp. 671-80.

Leopold, L.B., Langbein, W.B. (1962): 'The Concept of Entropy in Landscape Evolution', *U.S. Geological Survey Prof. Paper 500-A*, Washington.

Leopold, L.B., Wolman, M.G., and Miller, J.P. (1964): *Fluvial Process in Geomorphology* (San Francisco, Freeman and Co.).

Mandelbrot, B.B. (1967): 'How Long is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension', *Science, 156*, pp. 636-638.

Mandelbrot, B.B. (1983): *The Fractal Geometry of Nature* (San Francisco, Freeman).

Maritan, A., Rinaldo, A., Giacometti, A., Rigon, R., Rodriguez-Iturbe, I. (1996a): 'Scaling in River Networks', *Physical Review, E 53*, pp. 1501-1512.

Maritan, A., Colaiori, F., Flammini, A., Cieplak, M., Banavar, J.R. (1996b): 'Universality Classes of Optimal Channel Networks', *Science, 272*, pp. 984-986.

McMahon, T.A., and Bonner, J.T. (1983): *On Size and Life* (New York, Scientific American Library).

Metropolis, N., Rosembluth, M., Teller, M., and Teller, E. (1953): 'Equations of State Calculations by Fast Computing Machines', *J. Chemical Physics, 21*, pp. 1087-1096.

Montgomery, D.R., and Dietrich, W.E. (1988): 'Where do Channels Begin?', *Nature, 336*, pp. 232-234.

Montgomery, D.R., and Dietrich, W.E. (1992): 'Channel Initiation and the Problem of Landscape Scale', *Science, 255*, pp. 826-830.

Musu, I. (ed.) (1988): *Venezia Sostenibile: Suggestioni dal Futuro* (Bologna, Il Mulino).

Niemeyer, L., Pietronero, L., and Wiesmann, H.J. (1984): 'Fractal Dimension of Dielectric Breakdown', *Physical Review Letters, 5*, pp. 1033-1036.

Rigon, R., Rodriguez-Iturbe, I., Giacometti, A., Maritan, A., Tarboton, D., and Rinaldo, A. (1996): 'On Hack's Law', *Water Resources Research, 32*, pp. 3367-3374.

Rinaldo, A., Rodriguez-Iturbe, I., Rigon, R., Bras, R.L., Ijjasz-Vasquez, E., and Marani, A. (1992): 'Minimum Energy and Fractal Structures of Drainage Networks', *Water Resources Research, 28*, pp. 2183-2195.

Rinaldo, A., Rodriguez-Iturbe, I., Rigon, R., Ijjasz-Vasquez, E., and Bras, R.L. (1993): 'Self-organized Fractal River Networks', *Physical Review Letters, 70*, pp. 822-826.

Rinaldo, A., Dietrich, W.E., Vogel, G., Rigon, R., and Rodriguez-Iturbe, I. (1995): 'Geomorphic Signatures of Varying Climate, *Nature, 374* (April), pp. 632-636.

Rinaldo, A., Maritan, A., Flammini, A., Colaiori, F., Rigon, R., and Rodriguez-Iturbe, I. (1996): 'Thermodynamics of Fractal Networks', *Physical Review Letters, 76*, pp. 1501-1513.

Rinaldo, A., Rodriguez-Iturbe, I., and Rigon, R. (1998): 'Channel Networks', *Annual Review of Earth and Planetary Sciences, 26*, pp. 289-327.

Rodriguez-Iturbe, I., Ijjasz-Vasquez, E., Bras, R.L., and Tarboton, D.G. (1992a): 'Power-law Distributions of Mass and Energy in River Basins', *Water Resources Research, 28*, pp. 988-993.

Rodriguez-Iturbe, I., Rinaldo, A., Rigon, R., Bras, R.L., and Ijjasz-Vasquez, E. (1992b): 'Energy Dissipation, Runoff Production and the Three Dimensional Structure of Channel Networks, *Water Resources Research, 28*, pp. 1095-1103.

Rodriguez-Iturbe, I., Rinaldo, A., Rigon, R., Bras, R.L., and Ijjasz-Vasquez, E. (1992c): 'Fractal Structures as Least Energy Patterns: the Case of River Networks', *Geophysical Research Letters, 19*, pp. 889-892.

Rodriguez-Iturbe, I., and Rinaldo, A. (1997): *Fractal River Basins: Chance and Self-Organization* (New York, Cambridge University Press).

Sander, L.M. (19??): 'Fractal Growth', *Scientific American, 256*, pp. 94-100.

Shreve, R.L. (1966): 'Statistical Law of Stream Numbers', *J. Geology, 74*, pp. 17-37.

Shreve, R.L. (1967): 'Infinite Topologically Random Channel Networks', *J. Geology, 77*, pp. 397-414.

Stevens, P.S. (1974): *Patterns in Nature* (Boston, Little, Brown and Co.).

Tarboton, D.G., Bras, R.L, Rodriguez-Iturbe, I. (1989): 'Scaling and Elevation in River Networks', *Water Resources Research, 25*, pp. 2037-2051.

Wilson, E.O. (1998): *Consilience: the Unity of Knowledge* (New York, Knopf).

CRITICALITY AND SELF-ORGANISATION IN NATURAL PHENOMENA

LUCIANO PIETRONERO

1. The Simple and the Complex: Criticality and Scale Invariance

The traditional approach in physics has been to consider the simplest possible system and to study it in great detail. In this way, general laws have been derived which apply from the scale of the atomic nucleus to that of galaxies. This is the reductionist approach and it has been very successful in a broad range of situations.

However, it will be easily grasped, for example, that the properties of a living cell cannot be derived from the properties of the atoms from which the cell is built. The elements of the structure interact and lead to complex patterns and behaviour which have little to do with the properties of their isolated components. We may think of this as being the 'architecture' of matter and natural phenomena. It depends on the properties of the individual 'bricks' but it then develops its own characteristics and fundamental laws which cannot be related to those of the individual elements [1].

Apart from biology, which is extremely complex, physics already offers us a variety of complex structures such as trees, coastlines, clouds, lightning, turbulence in fluids, and the distribution of galaxies in the universe [2-5]. In these cases, complexity, for the most part, is represented in terms of scale-invariance, a property which emerges spontaneously from the nonlinear, dissipative dynamics of those systems.

Statistical physics is undergoing an important evolution [6]. The introduction of new ideas, based on fractal geometry and scaling, together with the study of dissipative processes with properties of self-organisation, provides us with the possibility of describing complex systems in terms of their unifying concepts and universality classes.

The physics of scale-invariant and complex systems is a novel field

which includes topics from several disciplines ranging from condensed matter physics to geology, biology, astrophysics and economics. This broad interdisciplinary approach corresponds to the fact that these new ideas allow us to look at natural phenomena in a radically new and original way, perhaps leading on independently to unifying concepts on the detailed structure of systems. The objective is the study of complex, scale-invariant structures which appear both in space and time in a vast variety of natural phenomena. New types of collective behaviour arise and their comprehension is one of the most challenging areas of modern statistical physics.

2. Fractal Structures and Self-organisation

Activity in this field has produced a co-operative effort in numerical simulations, analytical, and experimental work, which takes place at the following three levels.

(i) *The Mathematical or Geometrical Level*

This consists in applying the methods of fractal geometry to new areas in order to obtain new insights into important unresolved problems and contribute to an improved understanding of the whole. This allows us to bring within scientific areas many phenomena characterised by intrinsic irregularities which were previously neglected because of the lack of an appropriate framework for their mathematical description. The main examples of this type can be found in geophysical [7] and astrophysical data [8-9]. In particular, our work on the scaling properties of galaxy correlations represents one of the most challenging subjects in the field and has led to the much debated question of the fractal versus homogeneous universe at large scale [10].

(ii) *The Development of Physical Models: the Active Principles for the Generation of Fractal Structures.*

Computer simulations represent an essential method in the physics of complex and scale-invariant systems. A large number of models have been introduced to focus on specific physical mechanisms which can lead spontaneously to fractal structures. Here we list those which, in our opinion, represent the active principles for processes that generate scale-invariant properties based on physical processes. In refs. [3-5] one can find extensive descriptions of all these models:

Diffusion Limited Aggregation (DLA, 1981)
Dielectric Breakdown Model (DBM, 1984)
These models are the prototypes of the so-called Laplacian fractals in which an iteration process based on Laplace equation leads spontaneously to very complex fractal structures.
Cluster-Cluster Aggregation (Cl-Cl, 1983)
Invasion Percolation (IP, 1983)
The Sandpile Model (1987)
(The concept of self-organisation is common to all the models referred to here but it has been especially emphasised in relation to the sandpile model).
The Kardar-Parisi-Zhang Model of Surface Growth (KPZ, 1986)
The Bak-Sneppen Model (BS, 1993)

In addition to these simplified models, we know that fractal structures are naturally generated in fluid turbulence, as described by Navier-Stokes equations, as the fractional portion of space in which dissipation actually occurs. In addition, studies of gravitational instabilities suggest that gravity with random initial conditions may be enough to generate fractal clustering. However, the connections between the two important problems of turbulence and gravitational clustering, on the one hand, and the simplified models listed above, on the other, are only indirect. Each phenomenon and model mentioned seems to belong to a different universality class.

(iii) *The Development of Theoretical Understanding*

At a phenomenological level, scaling theory, based on usual critical phenomena, has been successfully used [4]. This is essential for the rationalisation of the results of the computer simulations and experiments. This method allows us to identify the relations between different properties and to focus on the essential ones. With respect to usual equilibrium statistical mechanics, these systems are far from equilibrium and their dynamics are intrinsically irreversible. This situation does not seem to lead to any sort of ergodic theorem and the temporal dynamics have to be explicitly considered in the theory.

From this perspective we have developed a variety of theoretical tools of a novel type which allow us to describe theoretically the self-organised generation of fractal structures [5]. These are: The method of the Fixed Scale Transformation which, together with the concept of Scale Invariant Dynamics, elucidates many properties of DLA, DBM and cluster aggregation. We have also developed the technique of run time statistics which is

able to map a dynamic in quenched disorder into a stochastic one with a long memory. This permits the study of invasion percolation and the Bak and Sneppen model. In addition, the dynamically driven renormalisation group allows the identification of the scale invariant dynamics in sandpile and related models and the analytical computation of their properties. Lastly, we have recently devised a new method to address the nonperturbative region of KPZ dynamics.

3. Perspectives

The ideas we have outlined here have been developed mostly within the physical sciences but one can well speculate that they are inherently interdisciplinary and should also be relevant to a variety of other fields such as geophysics, biology, economics, finance, and the social sciences in general. One could give many examples of more or less successful applications in these other fields. Usually it is important to have a vast amount of data in order to check the various hypotheses and identify the eventual scale invariant properties. The data should then be reproducible and it should be possible to make specific predictions which can be tested.

If one adopts these requirements the possible applications are drastically narrowed, at least for the moment. However, these new concepts and perspectives can motivate the acquisition of more data of high quality and it is our hope that this meeting will be able to stimulate important and innovative steps in this direction.

REFERENCES

[1] P.W. Anderson, *Science*, 177, 393 (1972).

[2] B.B. Mandelbrot, *The Fractal Geometry of Nature* (W.H. Freeman, New York, 1982).

[3] C.J.G. Evertsz, H.O. Peitgen and R.F. Voss, *Fractal Geometry and Analysis* (World Scientific, Singapore, 1995).

[4] P. Meakin, *Fractal Growth Phenomena* (Cambridge Univ. Press, 1998).

[5] A. Erzan, L. Pietronero and A. Vespignani, *Rev. Mod. Phys.*, 67, 545 (1995).

[6] See also J.L. Lebowitz, this volume, pp. 321-325.

[7] V. Keilis-Borok, this volume, pp. 289-295.

[8] P.H. Coleman and L. Pietronero, *Physics Reports*, 213, 311 (1992).

[9] F. Sylos Labini, M. Montuori and L. Pietronero, *Physics Reports*, 293, 66 (1988).

[10] See for example the various contributions in the volume: *Critical Dialogues in Cosmology*, ed. by N.Turok (World Scientific, Singapore, 1997).

IV.

SCIENCE AND PUBLIC POLICY

BASIC SCIENCE FOR DEVELOPING COUNTRIES

MIGUEL A. VIRASORO

It is generally recognised that economic growth can be to a large extent directly related to technological innovation while the latter ultimately depends on science, technology and engineering. This is true even when we understand progress in its deepest sense as improvement in the human condition in its health, cultural, environmental, social and economic dimensions.

It has also been generally accepted that this causal link becomes crucial in developing countries. The latter will be able to leapfrog and catch up with the more advanced societies only if they can access the latest technologies and scientific advances.

1. The New Knowledge Revolution

Today we have at our disposal powerful new tools as diverse as mathematical modelling, biotechnology and satellite observations to seek and find new solutions to defeat world-wide old problems, such as poverty and disease, and to address new challenges such as those presented to us by the deterioration of the environment.

But more deeply, we are in the middle of a conceptual revolution that for the first time recognises *knowledge* as the most precious economic good, much more valuable than the usual natural resources that have dominated our definition of wealth up to very recently.

Knowledge distinguishes itself from other products of human activity by two features. The first one is that the use of knowledge by one person does not prevent in any way the use of that same piece of knowledge by another person. This is different from other kinds of good (consider food for instance). The second feature is that once some information becomes public it is difficult to prevent anyone from using it. For instance, once a theorem has been demonstrated and published anyone can use it in his/her

own research to produce new results (the same consequence applies when we refer to an experimental discovery or an explanation of an experimental fact). Economists call these properties non-rivalry and non-excludability and they recognise that goods of this kind require a special kind of treatment. A competitive perfect market for them requires complex regulations and does not necessarily lead to an efficient allocation.

In particular, for a non-rival good, the most efficient allocation is to make it available at no cost to as many people as possible; it is obvious that when more people know about a discovery, there will be more occasions to derive from it new consequences.

Unfortunately, this is not the whole story because economic thinking stresses the need to reward the creator of knowledge. We can here distinguish two extreme proposals:

1. Society recognises the property right of the intellectual creator to his own work and creates institutions and regulations with the purpose of allowing him to control the access to his work. This permits the creation of a market that then fixes the price of access. Thus a regime of Intellectual Property Rights (IPR) is installed.

2. Society defines the intellectual product as a public good and rewards the creator at the source.

The first mechanism works well, for instance, for the entertainment industry but it does not work equally well for scientific knowledge in general and is totally inadequate for Basic Science. Here excludability is near to impossible. There are IPRs in Science but today the author transfers them, *for free*, to the publishing companies. Furthermore, any individual has free access to the information in public libraries. Finally there is the concept of "fair use" which allows teachers and researchers to make their own photocopies without having to pay copyright fees.

Today, the obstacles to a world-wide dissemination of scientific knowledge lie exclusively in the relatively high cost of production and delivery of journals and books. This turns out to be dramatic for developing countries. Very few libraries in the South have complete collections even of the most important journals.

The information revolution that we are witnessing has two effects. On the one side it decreases the cost of dissemination to basically zero; on the other it almost nullifies the mild restrictions to accessibility legislated in the present IPR regime: once we download an article on a computer we can, by simply pushing a key, send copies of it to hundreds of colleagues. So the concept of "fair use" is at stake and in fact there are several projects to eliminate it. That is: the new regimes of Intellectual Property Rights being

discussed today both in Europe and in the US are much more strict than the old ones. This is the consequence of a terrible miscalculation. On the one hand it will not attain its purpose because results can be simply rephrased and there is no law that prevents that. On the other hand it will reduce considerably the positive impact that the new communications technologies could have on the developing countries. At a moment when researchers in those countries could finally electronically access the best libraries in the world, they will discover that they cannot do that because they cannot pay for the access rights.

There is a related dilemma when decisions have to be taken concerning basic science funding. The private sector is not happy with the idea of paying for something that could benefit a competitor. A private firm could support basic research if either it enjoys a monopolistic situation or it estimates the time lag between the basic research result and its application as being so short that it can reasonably expect to reap sufficient benefits from being the first to know the result (this is the case today with genetics research).

It is a theorem in economics dynamics that the private sector alone would strongly under-invest in basic research. This is due to the fact that, in any case, the social returns are larger than the benefits a private investor can appropriate. An analysis of social return to investments in academic research in the USA (which are either on basic science or share the same problems of appropriability) deserves quoting. In the period 1982-1985, 76 firms belonging to 7 industries identified new products that could not have been developed at all or would have suffered substantial delay in the absence of recent academic research. Those products accounted for about $24 billion of sales in 1985 alone. The mean time lag between the successful research result and the appearance of the product in the market was calculated to be 7 years. Taking into account total spending on R&D in the academic sector and other expenses needed to generate the final product (industrial R&D, plant and equipment and start-up activities), this study (Mansfield 1991-1992) estimates a social rate of return that exceeds 20%. This is very high and it is still an underestimate because it does not include, for instance, the educational value of research.

If we add that private rates of return come out typically a factor of 2 or 3 smaller than the rates of return to society, the moral is simple: the public sector should take the burden of the necessary investment in basic science research, either directly or indirectly (by tax incentives) leaving, if they wish to do so, the technological applications to the private sector.

Therefore as has been already suggested, basic science knowledge

should be considered a public good and the burden of financing academic and basic science should, to a large extent, be taken by the public sector.

One last point: for this to be effective all countries should contribute in a fair measure. We are effectively proposing the internationalisation of basic science.

2. The Public Perception of Science and Basic Science. The ICTP Perspective

By definition basic science research is intended to advance fundamental knowledge, irrespective of any foreseeable application. It is a historic fact that such speculative research has produced the farthest reaching conceptual revolutions which have paved the way to key technological changes. But it would be misleading to think that one can plan the course that sciences would take so that they end up in specific applications. In the Anglo-Saxon literature they talk of curiosity-driven research and in fact it should be seen as the realisation of that spiritual instinct that drives humanity to study nature and to seek the ultimate consequences of its own reasoning.

If the opportunity of investing in basic science is questioned in rich countries today, the dilemma becomes dramatic in developing countries where well-intentioned decision-makers may recognise the importance of basic sciences, but worry whether they can afford to invest in them given the depth of their immediate economic and social problems.

International aid agencies have been particularly sensitive to these arguments so that some of them (i.e. the last V Framework Programme of the European Commission) explicitly rule out any support to basic science.

However, the consequences for the developing world, especially the least developed countries, could be disastrous.

Here I am speaking on the basis of my own experience in Argentina but much more importantly on behalf of thousands of scientists from the Third World who bring their experience to our Centre every year. Let me now open a parenthesis to make a short presentation of this unique institution.

2.a *The Abdus Saiam International Centre for Theoretical Physics*

This Centre is one of the few institutions in the world dedicated to the ideal of promoting and supporting basic science in developing countries. It is an international centre under the umbrella of the IAEA-UNESCO, conceived and created by Abdus Salam and transformed into reality with the help of another visionary, Paolo Budinich, and the generosity of the Italian Government. Its institutional objectives are:

Institutional Objectives of ICTP

To help in fostering the growth of advanced studies and research in physical and mathematical sciences, especially in developing countries.

To provide an international forum for scientific contacts between scientists from all countries.

To provide facilities to conduct original research to its visitors, associates and fellows principally from developing countries.

Today the reality of the Centre can be summarised in a few statistics: about 2500 physicists from the Third World and 1500 from the rich, developed countries visit the ICTP every year for periods ranging from one week to several months. On the average one can encounter 400 visitors every day during the year. They come to do research using one of the largest libraries in Europe, a network of almost 400 computers and the possibility of collaborating scientifically with one of our 16 scientific staff plus the almost 50 researchers from SISSA (an Italian School of excellence) plus of course the other 400 visitors.

We organise every year about 40 activities ranging from 5 week Colleges to 2 to 3 day mini-conferences on subjects that cover all fields in mathematics and physics and even other disciplines provided they use methods borrowed from physics and/or mathematics. Some examples are: School on Complex Disordered Systems, School on the Mathematics of Economics, Algebraic Geometry, Strongly Correlated Electrons, String Theory etc. One of our programmes that has been very praised and copied is the Associate Scheme. Through it at any moment 600 scientists working in developing countries represent our external faculty and enjoy special privileges such as the possibility of visiting the Centre (3 times in 6 years) alone or accompanied by a student or sending his/her student to one of our training activities.

In every region of the world one can find Associates and Former Associates of the Centre. In many cases this elite, which recognises the ICTP as its second alma mater, plays an important role in shaping the scientific policy of their country. There are Associates that have become Ministers or Secretaries in Scientific Affairs. A few have been named in still higher political positions. A large fraction of the Presidents of the Physical Societies are also former Associates of the ICTP.

In the least developed countries we count 17 Affiliated Centres that

receive every year a small grant from the ICTP to build up their facilities. In addition we sponsor and partially fund South-South Networks and around 40 conferences in developing countries.

It is the experience accumulated along its 35 years of existence that I would like to bring here to bear on the question.

2.b *Why Basic Science in Developing Countries?*

The first reason, and perhaps one of the most important, into aid scientists from developing countries to enter the mainstream of research in the last frontiers of knowledge, is the deep moral conviction that in the great adventure of exploring nature, no cultural group, no nation should be left aside. This is an ethical-political issue. Ethical because it has to do with equal dignity for all the people of the world. Political because it stresses the ultimate unity of our planet. Some criticism always comes to us from the rich countries and concerns waste of resources. This is an extremely narrow point of view. The figure of Abdus Salam is a case to study. Should he have quelled his curiosity that pushed him to seek the fundamental laws of nature so as to dedicate his life to alleviate the sufferings of his fellows? I am convinced that a crushing majority of Pakistanis think differently. Even if they have derived no material benefits from Salam's discoveries they are proud of him and that pride plays an invaluable role in any society.

I would like to suggest that participation in scientific research projects should be considered as one of the fundamental human rights of each individual, as the natural continuation of the already recognised right to higher education and as a part of the cultural rights indispensable for the dignity and the free development of the personality as discussed in Articles 22 and 26 of the Universal Declaration on Human Rights.

One cannot forget the real suffering faced by millions of people in the developing world; nor can anyone suggest that nations should postpone those urgent immediate steps–which are necessary to improve the economic and social well-being of their people. However, if the effort to solve immediate problems prevents the building up of indigenous know-how, then the problems of today will be the problems of tomorrow. More specifically, and as an example of possible plans of action, when international agencies finance goal-oriented projects, they should simultaneously help to build up a solid scientific establishment so that knowledge is effectively transferred.

Applied research is of direct relevance and importance, but it requires:

— critical mass in several basic sciences;
— continuity of management and funding;
— either international collaboration or local expertise.

A good foundation in basic sciences is therefore an indispensable ingredient. Basic scientists will educate young researchers who will profit from applied research projects. In many cases in developing countries, basic scientists have gone into applied research programmes. Last but not least, basic science, being from the start international and academic in character, is subject to that tight and transparent quality control that every society is entitled to expect.

This problem of quality control has been generally overlooked though the consequences of such an attitude could seriously jeopardise valid efforts to build up a scientific infrastructure. In general, the problem can be rephrased in the following way: how can non-experts (the public in general, government officials) judge the quality of an expert? Similarly, in a world where solutions to problems increasingly involve scientific knowledge, how can public officials and citizens determine which experts to trust when there are competing opinions? Issues that call for a science-based solution normally elicit competing proposals that rely on powerful lobbies to gain support – and ultimately funding – for their different strategies. Without access to university-based scientists capable of assessing the merits of each proposal, public officials and citizens alike will be at the mercy of those who have vested interests in the proposals that they are presenting.

As normally happens in developed countries, the practical solution relies on strong academic institutions, and, more specifically, on the fact that scientists working in basic research can be objectively evaluated. This is because of the free dissemination of basic science knowledge and because the open problems are shared and known by all those working in the same discipline. Such an assessment will never be exact but even if approximate it can be extremely useful.

For applied research this information is more difficult to obtain. First of all, because publications are less relevant. Second, because these experts are not necessarily organised in disciplines. Third, because, by definition, the problems they address are unique events that do not reproduce exactly.

The North has long called on its basic scientists to serve as impartial judges for assessing controversial science and technology issues of vital national concern. An interesting relatively recent example was the calling of Richard Feynman, Nobel Laureate for theoretical physics, to assess the causes of the Challenger shuttle disaster. The story is known but it shows how society used his name and impeccable credentials of impartiality to recreate the trust that was reduced between NASA and the public.

In short, well-trained basic scientists have the power to bridge the gap between scientific experts and the public in ways that make science-based development possible.

3. International Co-Operation in Basic Research Capacity Building

In addition to the ethical reasons we referred to at the beginning of the last section there are other more practical reasons that suggest that it is in the interest of rich countries to help developing countries to build capacity in the basic sciences. With globalisation, problems are also becoming global. Environment deterioration is one example but there are others like, for instance, natural resources depletion or insufficient health care prevention programmes. We have already argued that capacity building in the basic sciences is an important preliminary step towards creating networks of experts capable of addressing these global programmes.

On 15-16 June 1995, an "International Conference on Donor Support to Development-Oriented Research in the Basic Sciences" was held at Uppsala University, Sweden. Among the convenors there were several agencies with pluriannual experience in basic science co-operation. They included:

IPPS - International Programme for Physical Sciences, Uppsala, Sweden
IFS - International Foundation for Science, Uppsala, Sweden
TWAS - Third World Academv of Sciences. Trieste Italy
ICGEB - International Centre for Genetic Engineering and Biotechnology, Trieste, Italy and New Delhi, India
SIDA - Swedish International Development Cooperation Agency, Sweden
ORSTOM, Bondy, France

The conclusions are worth quoting:

— a foundation in the basic sciences is essential for all research in the applied sciences and for long-term development;
— adequate funding for the basic sciences from domestic support and external aid programmes is necessary.

and the measures proposed were:

1) Capacity building in the basic sciences.
2) Support for researck and higher education in the basic sciences. In particular donor support to applied projects should include grants to research and higher education in the basic sciences.

The World Bank has recently introduced its own programme and proposes the Millennium Centres project with special emphasis on excellence.

I am confident that the time has come when we do not have to argue anymore whether basic research is a luxury that developing countries can not afford.

4. Scientists in the Developing World

As Abdus Salam has often underlined, it is reasonable and fair to expect that the scientific community should pay special attention to the problems that afflict society and hamper its development. Scientists who are continuously measuring themselves in the international arena, contributing towards the great adventure of pushing backwards the frontiers of the unknown, bring with them a fresh attitude of intellectual rigour, attention to the facts and a precise language that are key ingredients in the search for solutions to these problems.

The Abdus Salam Centre has begun to follow the careers of those who have passed through the Centre as Associates, long-term visitors or post-docs. The startling and rewarding result is that many of them have been called to assume positions of high responsibility in their respective governments.

The ICTP is proud of this fact because it shows a high degree of sensitivity towards social problems. In some cases this does not happen. I have heard many arguments to justify failure. Two of them are relevant for our future action:

1) The education received does not provide the necessary tools for analysing the complexities of real world problems.

2) The imprecision of the data, the lack of knowledge of all the variables that parameterise a problem, and the "dirty" character of the latter prevents a scientific treatment of it.

Both these arguments stem from the same root and point towards the same remedy. In tackling problems which come from the real world, the key difficulty lies in their re-formulation in terms that are conducive to scientific treatment. This means extracting the relevant variables, identifying their interactions and in some cases formulating goals in mathematical terms. In this way one constructs first approximations that can at best be under control, but at least can be refined if proved inadequate.

Recently both mathematics and theoretical physics have made progress in tackling complexity and uncertainty. "Complex systems" has become an active area of research. There has been a flourishing of new curricula on "mathematical modelling", "decisions under uncertainty", "statistical assessment and management of risk" and many others. Of course, this progress is closely connected to the explosive introduction of the Computer as an instrument to analyse models.

The ICTP has followed this trend and has organised many courses dealing with these subjects. To name just a few of them:

Statistical methods applied to medicine

Mathematics of economics (including one week on the economics of public goods directed by Partha Dasgupta)
Neuro-informatics
Mathematical ecology
Statistical physics and computer science
Modelling the real world: an introduction to industrial mathematics
Comparative assessment of different energy sources
Weather and climate dynamics

On the research side we have recently created a new group on the weather-climate system, one of the most complex non-linear systems that has ever been studied thoroughly using numerical modelling, statistical methods, scale analysis, boundary layer theory, asymptotic expansions and their matching. There has also been a flurry of new activities using statistical methods applied to the modelling of complex systems: modelling economic systems, game theory, the brain, glasses, phase transition concepts in computer science.

We want to intensify our efforts in these promising directions. We thus expect to train a new kind of scientist who is even more prepared to deal with the difficult reality that he has to face everyday in his own country.

RUSSIAN SCIENCE: DOWN THE UPWARD STAIRCASE

VLADIMIR E. FORTOV and LEVAN MINDELI

The Russian science and technology system is one of those areas of activity where a difficult and even critical situation has emerged during the last few years. The shocking nature of reforms started in 1992, lack of an elaborate conception of socio-economic transformations, and distinct ideas concerning the role of S&T and innovation in the period of global transition from plan-based to market-oriented mechanisms of economy management, all originated a whole range of negative trends in the R&D sector which are now being experienced by the once powerful S&T complex of Russia.

The decisive characteristics of the present situation are as follows: multiple reduction of R&D funding from all sources; a considerable decrease in the number of researchers, followed by a decline of their living standards as well as in the material and technical provision of R&D activities, and other negative processes; decline of demand for S&T output in the domestic international market; and destruction of the management basis of science and technology as well as the available material and technical facilities.

At the same time, it should not be ignored that the institutional structure of the Russian science and technology sector, its internal correlation and mechanisms of operation, had been mainly formed long before the start of economic reforms and were far from always contributing to an effective integration of science in the sharply changed circumstances of economic activities. R&D institutions and researchers themselves, confronted with the new and unfamiliar environment, tried to adapt to the situation. However, this adaptation took place in the absence of a well-timed and adequate to the changes reaction on the part of the government and hence was very painful. Essentially, in this very important process for Russia, the government turned away from assistance to science, whereas the R&D executives and elite turned out to be incapable of convincing the political leaders of

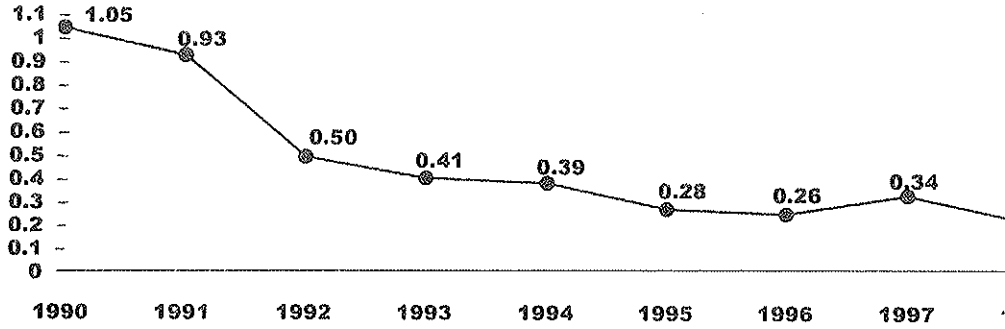


Fig. 1. Allocations by the Federal budget to basic research and promotion of S&T progress (*as a per cent of GDP*).

the need to protect science in the crisis period and proposed no efficient mechanisms to deal with the dramatic situation.

The most critical circumstance for scientific potential is the reduction of government appropriations for R&D. During 1991-98, the budget funding of R&D at comparable prices decreased almost tenfold. The ratio of R&D expenditure to GDP decreased fivefold for that period (fig. 1). By the level of this indicator, Russia has approached the group of countries whose scientific potentials are hundreds of times less than that of Russia, such as New Zealand, Spain, and Portugal (fig. 2).

For the period 1991-98, expenditure per head of researcher decreased in the Russian Federation almost fivefold, something which has made Russia's level 25-30-fold as low as that of advanced industrial countries.

The negative trend in the values of annual government allocations to R&D, formed during those years, is further aggravated by the non-execution of budget obligations (fig. 3). The R&D budget for 1995 was fulfilled only 73.4 per cent; in 1996 R&D was actually financed at 62.3 per cent of the planned level. In 1997, the situation in R&D funding improved to some extent in comparison with the preceding years 1992-96 (the budget was fulfilled at a level of 83.2 per cent). The expenditure on R&D from the Federal budget was 2.2 times as much as in 1996. However, last year the growth trend in R&D funding could not be maintained. In 1998, the R&D sector received only 57.3 per cent of the initial budget allocations. In the budget for the current year, the R&D item maintains the figure on a level with 1998. However, the August 1998 crisis sharply decreased (threefold) the real value of this amount.

The reduction in the scale of R&D funding in the transition period was

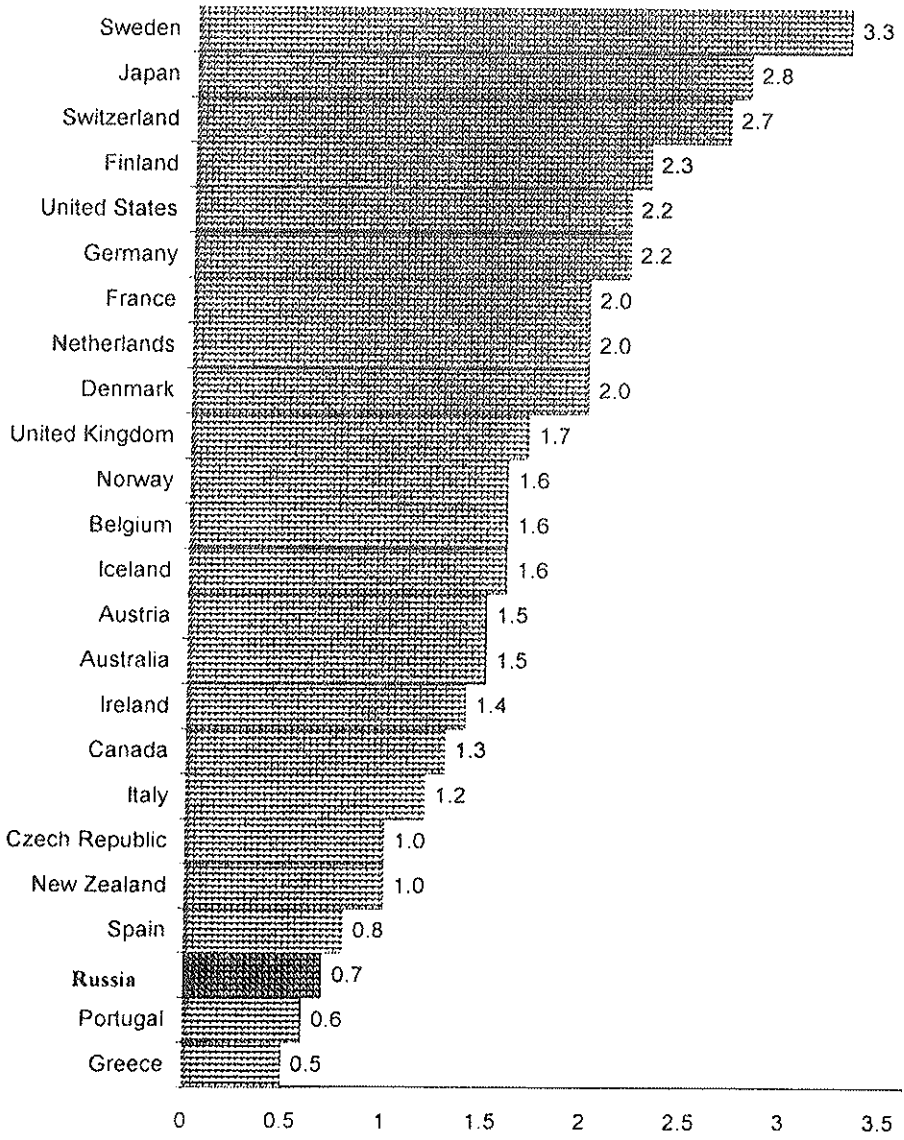


Fig. 2. Gross domestic expenditure on R&D in Russia and OECD countries (*as a per cent of GDP*).

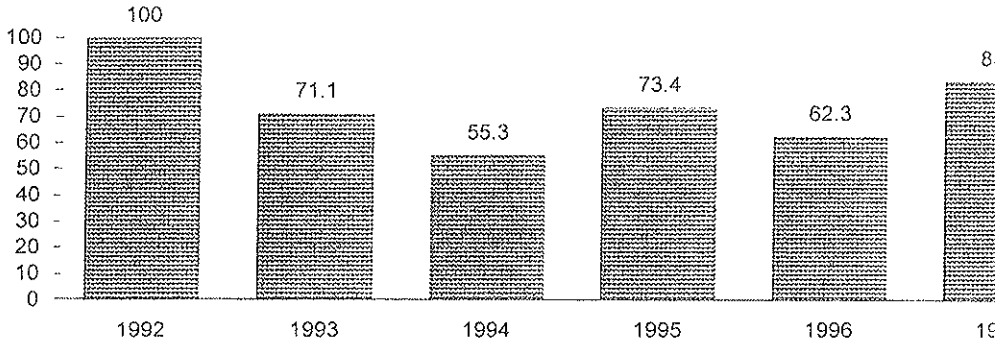


Fig. 3. Fulfilment of R&D budget (*per cent*).

accompanied by certain shifts in the composition of the sources of these funds. Under the conditions of the planned economy in the pre-reform period, the Russian R&D sector was functioning within a framework of the centralised accumulation and distribution of financial resources through the state budget. Given that firms' funds under those conditions were essentially part of redistributed budget resources, their actual part in R&D funding was very insignificant, more than 4 per cent of its total.

For the purposes of funding R&D and compensating for expenditure related to development and the introduction of new products and technological processes in industrial ministries and departments of the former USSR, the United Fund for S&T Development (UFSTP) was established. This fund was made up largely of deductions from the planned profits of research and production organisations and industrial enterprises according to a norm fixed for a particular Ministry, as a per cent of net (in some cases commodity) output, for five years with the divisions by years. Allocations to the Fund were made monthly, as a rule by all enterprises with a planned profit. The UFSTP financed two-thirds of planned works.

At the beginning of the reforms the self-regulating function of the market was exaggerated.

The prevailing point of view on the part of public opinion was that in a market economy applied R&D should be performed on the basis of self-repayment and according to demand from companies and enterprises of both the public and private sectors, and would progress in line with the impact of the market on the former industrial sector which would easily be transformed into company-based R&D. However, practice has demonstrated that the R&D system formed under conditions of the administrative

planned economy cannot adapt automatically to the market-oriented economic methods based on the relationship between demand and supply in S&T output. The situation is further aggravated by the extremely unfavourable investment climate in the country, caused by the general crisis, which impedes innovative activity and makes the R&D sector unattractive for free capital.

To prevent the destruction of national S&T potential and to some extent to compensate for the dramatic reduction in the R&D expenditure of industrial enterprises, non-budget foundations were established in the country to finance sectorial and intersectoral R&D. Their resources are formed from voluntary deductions from these enterprises and equal 1.5 per cent of their sales. 25 per cent of funds accumulated by these foundations go to the Russian Foundation for Technological Development.

The emergence of a network of budget and non-budget R&D foundations in Russia has involved important changes in the mechanisms of monetary funds in the implementation of S&T and innovation projects. The existence of these foundations has brought about:

— a strengthening of the selective financing of S&T and of innovative projects through the introduction of principles of competition;
— an enhancing of the sources of funds for S&T and innovative projects;
— a stimulating of the initiative of researchers and innovators in the search for funds to finance their work.

In the process of transition to a market economy, the role of foreign sources of funds for R&D intensified at a sufficient level. Their share increased from an almost zero level on the eve of the 1990s to 7.4 per cent in 1997, including a figure of up to 30-50 per cent in basic research.

In general, it can be declared that, despite the remaining financial constraints, the centralised system of R&D funding is gradually being replaced by a diversity of sources of funds, and incentives for private investment in R&D are gaining in importance.

The emergence and development of the system of multichannel financing of R&D, on the one hand, and the attempts of R&D institutions themselves to adapt to the changing conditions, on the other, are factors which influence the scale and trends of development of particular R&D sectors. Thus, although the business enterprise sector is now, as before, the largest of all sectors in R&D (its share in the structure of R&D expenditure is two-thirds), the nature of R&D performed within it has changed: the proportion of long-term applied research has decreased to the benefit of R&D aimed at the satisfaction of current industrial demand. This change in the structure completely corresponds to the drastic reduction in large long-term invest-

ment projects and the excessive weight of projects capable of giving a quick commercial profit within the strategies of industrial enterprises.

The proportion of basic research within the business enterprise sector of R&D has remained practically stable during the past decade; at the same time in the government sector it has increased. As regards R&D units of industrial enterprises, here there has been a redistribution of the portfolio of orders in favour of development.

The share of the government sector in the structure of R&D expenditure has noticeably increased. This share is more than twice as much as the OECD average, a development caused by the important role of academy institutes active in this sector in carnying out R&D in Russia.

The contribution of the higher education sector to R&D efforts is insignificant. Its share in R&D expenditure was at the beginning of 1998 only 5.4 per cent. This is a specific feature of the R&D system in Russia, something which makes it different from the Western countries where the role of universities is very important.

Desperately needing additional financial resources, many R&D institutions have enlarged the proportion of new activities in the total value of their work. Although this tendency emerged in the late 1980s in connection with the mass-scale transition of R&D institutions to a profit-and-loss basis and self-financing, in the period of reforms rendering various services (information, marketing, printing, etc.), renting premises and equipment, and in some cases manufacturing products, have become for many research institutes a substantial, if not the principal, source of income. In comparison with the pre-reform period, the share of unconventional activities in the total value of R&D institutions' work has increased twofold. Taking into account the shortcomings of the taxation system and the practice of accounts, the significance of this indicator may increase further in a significant way.

Thus the inertial development of science, combined with the deep financial crisis, entailed deformations in the structure of the R&D potential and a deterioration of its qualitative characteristics.

In comparison with the year 1990, which was favourable for S&T activities, the population of R&D personnel has reduced by more than twofold. This is considerably more than in such budget-funded areas as education and health services (fig. 4).

It is noteworthy that the Russian R&D sector still remains one of the world's leaders in terms of levels of employment. By this indicator, Russia is now, as before, ahead of many industrial countries (table 1). This is an additional proof of the fact that human resources are the mainstay of Russian

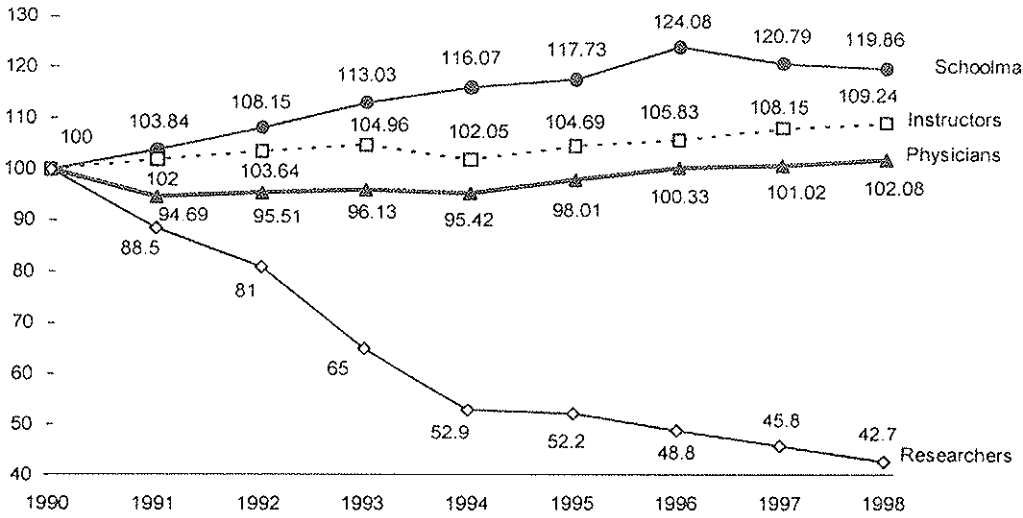


Fig. 4. Trends in qualified manpower (*per cent*).

science and concern about them must play an important part in state S&T policy.

People leaving R&D institutions considerably outnumber those who join them. Meanwhile in the structure of employment, personnel reduction takes place mainly at the expense of researchers and technicians (fig. 5).

The process of R&D personnel reduction has affected various categories in different ways. During the first stage of reforms, dismissal to a greater extent affected technicians, supporting staff and other auxiliary personnel than other categories. This is explained by attempts to preserve the research teams as the key element of R&D human resources, and to reduce overheads. At the same time, the decrease in the number of technicians, laboratory assistants, equipment operators, and personnel of testing and experimental production units inevitably has affected the working conditions and the efficiency of researchers (who often have to combine research with technical functions) as well as the condition of instruments and equipment. By 1995-96 the number of technicians stabilised at 9-10 per cent of the total R&D workforce, and that of supporting staff and other auxiliary personnel at 41-43 per cent; whereas the proportion of researchers is now less than half. Given the further decrease in the number of researchers, this testifies to the fact that in certain research institutes proper scientific activities are gradually fading away.

Table 1. *Researchers by country** (hierarchical rating by absolute number).

Country	Researchers (thousand)	Researchers per 10,000 head of population	Population (million)
United States	1385.0	52.4	264.0
Japan	660.0	52.8	125.0
China	430.6	3.5	1231.0
Russia	**428.6**	**29.1**	**147.1**
France	153.0	26.3	58.1
United Kingdom	142.0	24.5	58.0
India	110.0	1.2	928.0
Korea	107.3	24.0	44.7
Italy	74.0	12.9	57.3
Canada	68.1	22.6	30.1
Australia	53.0	29.4	18.0
Spain	47.6	12.1	39.3
Netherlands	31.0	19.9	15.6
Sweden	30.1	34.2	8.8

* In this case it is necessary to take into account that not all the persons formally employed in R&D are really working in this sector.

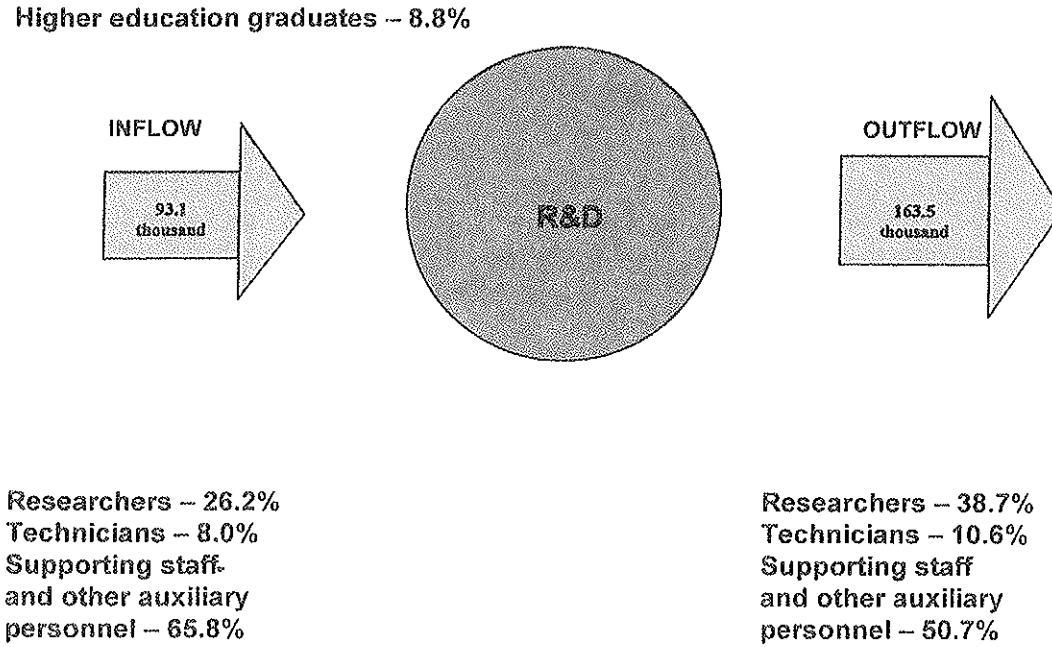


Fig. 5. Flows of R&D personnel.

The reduction of R&D personnel is immediately related to dynamic trends in the labour market and is on the whole spontaneous in the context of a lack of government regulation. Currently, the main factor is the voluntary outflow of employees from the R&D sector: in 1997 it was 58.9 per cent of the total outflow of personnel from this sector against 16.3 per cent dismissed for redundancy. In the first place, this is the so-called internal brain drain, i.e., departures of highly qualified scientists and specialists to the business sector. Opportunities that had emerged there enabled many of them to find, without much effort, well-paid and promising jobs; as a result, many managers of banks, investment and industrial companies, joint ventures, and other large business entities are ex-R&D professionals. This redistribution of highly qualified manpower certainly has contributed to the rise of sectors of the market economy that are completely new for Russia, but at the same time it has been a serious blow for Russian science. Meanwhile, employees with less qualification, confronted with the difficulty of employment, have often come back to lower paid positions at budget-funded institutions, joining supporting staff and other auxiliary personnel.

In 1998, such employees were two-thirds of the personnel inflow in the R&D sector.

Statistics of salary rates in the R&D sector demonstrate that, during the last year, there has been some success in attempts to decrease the gap between pay rates in this sector and the average wage rate in the national economy in general (fig. 6). Nevertheless, the rates of R&D salaries are still low. This failing, which has lasted a long time, not only has injured the reproductive processes in R&D personnel but has also raised the danger of an irreversible destruction of the scientific potential and the culture of R&D activities in the country.

Low salary rates have decreased the attractiveness, and entailed a

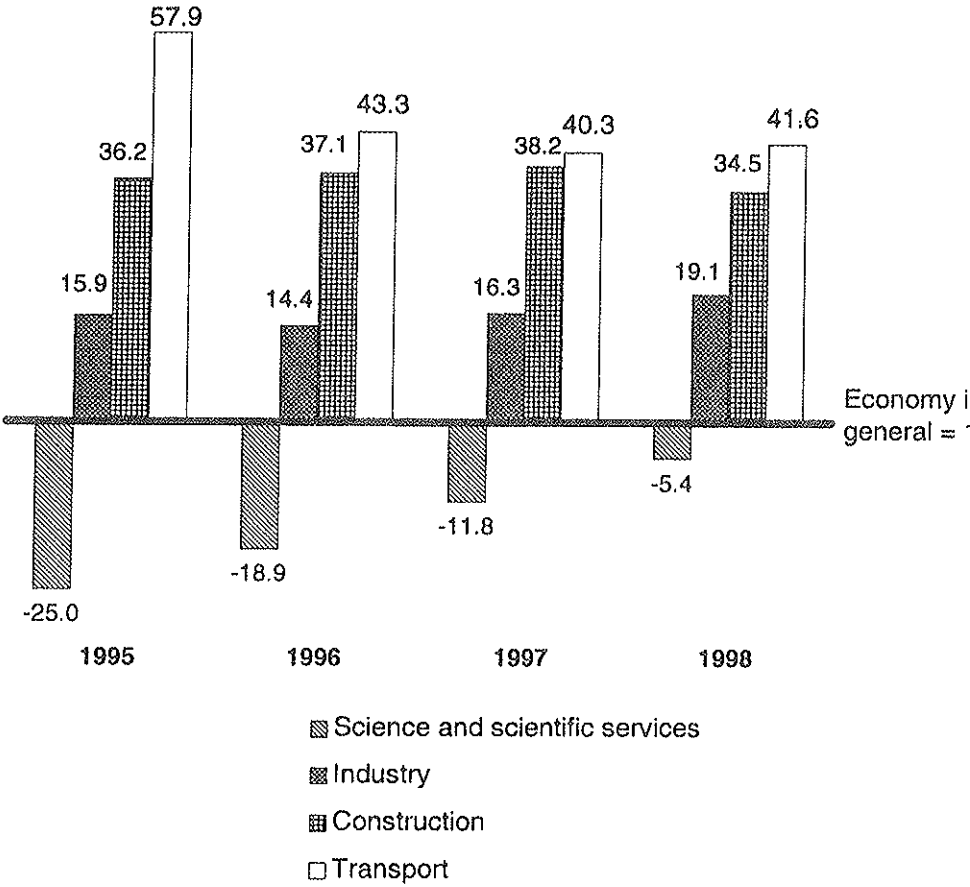


Fig. 6. Average monthly wage and salary rates (*per cent*).

decline, of scientific occupations. For this reason, many talented researchers have changed occupation, and in addition the inflow of graduates from higher education institutions into the R&D sector has dramatically decreased.

Statistics testify to a gradual decrease in the proportion of researchers aged below 30, as well of the age group from 30 to 49 years. At the same time, the proportion of older researchers (50 to 59 and over 60 years of age) is accordingly increasing. The ageing trend is especially distinct in S&T organisations of non-productive sectors, where, in a year and a half, the proportion of researchers aged below 50 decreased from 56.8 to 53.3 per cent.

A comparison of data for 1988 and the findings of a survey of 1996 demonstrates that the proportion of researchers aged below 50 decreased from 78 to 55 per cent, whereas the proportion of those aged from 50 to 60 increased from 17 to 31 per cent, and that of those aged over 60 increased from 5 to 13 per cent.

Thus there has been a considerable increase in the proportion of older researchers. For example, in the Russian Academy of Sciences the average age of academicians has reached that of 70.

The trends in the human resources of Russian R&D are certainly influenced by an international migration of scientists in the form of emigration as well as departures for temporary work abroad. According to estimates, the number of emigrating scientists, comprising all categories of employees, does not exceed 2 thousand a year. The departures of R&D employees for permanent residence abroad are largely driven by ethnic factors, and this sometimes involves the best qualified professionals.

It should be noted that emigration processes have entailed quite appreciable losses in some areas of the Russian R&D sector. According to V. Arnold, an Academician of the Russian Academy of Sciences, this is the situation in Russian mathematics because many gifted mathematicians have left Russia.

Russia's participation in international S&T links has contributed to the departure of scientists for temporary work abroad. Whereas in 1991-92, 1.7 thousand researchers of the Russian Academy of Sciences were abroad under long-term contracts (2.8 per cent of RAS total R&D employment), by 1993 they were 2.6 thousand (3.1 per cent).

A more comprehensive representation of the scale and structure of Russian scientists' departure for work abroad is provided by a survey conducted by the Centre for Science Research and Statistics.

According to this survey, in 1996 4,084 Russian researchers from 280 research institutes and universities were temporarily working abroad, which is 2.6 per cent of the total R&D employment in these institutions. Academy

institutes numbered 2,727 researchers of this category (6.4 per cent of the total employment of researchers in these institutions). In R&D institutions of particular sectors this category of researchers amounted to 501 (0.5 per cent), and in higher education institutions it was 856 (7.8 per cent). Accordingly, it is possible to recognise that academy institutes, although leading by the number of employees working abroad, are nevertheless to some extent inferior to higher education institutions when it comes to the intensity of involvement of researchers in this process.

More than half of Russian researchers (52.9 per cent) who had left the country were working in the United States, Germany, and France, and another 13.5 per for each country were in the United Kingdom, Italy, and Japan. In this way, the six above countries accounted for two-thirds of all Russian researchers working abroad, whereas the rest of the countries had only one-third.

The majority of Russian scientists temporarily working abroad consisted of professionals in the natural sciences, especially mathematicians, physicists, and biologists. Next to them, at a great distance, were researchers in engineering and humanities. The proportion of researchers in social sciences and medicine among those working abroad was insignificant (5.2 and 4.8 per cent, respectively), although it was slightly above the proportion of those employed in these fields of science in Russia as the whole.

The situation concerning the material and technical provision of R&D is also extremely unfavourable. The material and technical base of R&D is ageing. The share of machines and equipment in R&D fixed assets has been reduced more than twofold. The R&D infrastructure is going to pieces. The influx of information in Russia is now 3.5-fold less than in 1991 and twofold less than the necessary minimum level of information that the country should receive.

At the same time, it should be remembered that even in the extremely hard times of World War II, during the Battle of Stalingrad in 1942, subscription to foreign journals was not interrupted. New periodicals arrived even in blockaded Leningrad, including publications from Germany and Italy. The publication of scientific literature during the war was growing annually. In 1942 the Academy of Sciences of the USSR published 350 books totalling 2,300 printer's sheets; in 1944, 496 books totalling 3,750 printer's sheets were sold; and in 1945 the sales were equal to 570 books totalling 5,600 printer's sheets. The publication of scientific journals was maintained at the pre-war level.

In present-day Russia, as a result of output decline and the insufficient funding of S&T, the volume of current R&D projects is decreasing; business entities have to a considerable extent lost interest in re-equipping the

Table 2. *Patent applications filed with Rospatent.*

	1993	1994	1995	1996	1997
Patent applications submitted to Rospatent	32216	23081	22202	23211	19992
Of which from resident applicants	28478	19482	17551	18014	15106

production and technological base. At the same time, indicators of levels of development of inventive activities as well as patenting and licensing are decreasing; for example, such an important indicator as the number of applications filed with the Russian Patent Office (Rospatent) (table 2).

During the last four years, the number of invention applications submitted by foreign applicants has been relatively stable at the level of about 13.5 per cent. In comparison with patent offices of advanced industrial countries, where the proportion of foreign applications varies around 50 per cent, the low proportion of foreign applications submitted to Rospatent may characterise the lack of economic interest in patenting advanced technologies in Russia and in the protection of exports on the part of foreign countries.

The period of economic transformation in Russia, associated with a drastic decline of output and solvent demand, has also been characterised by a reduction in innovative activity. In the late 1980s this indicator of industrial enterprises of the former USSR varied within the range 60-70 per cent. With the beginning of the reforms, its value decreased three- to four-fold, with the ratio of innovating enterprises to the total number of industrial enterprises in 1992 equal to 16.3 per cent, in 1993 to 17.3 per cent, and in 1994 to 19.5 per cent. Such a slashing decline was then explained with reference to the difficulties of the transition period. It was supposed that the processes of reforming the economy would give an impetus to an increase in production, the intensification of S&T development, and the replenishment of the budget. However, nothing of this kind happened. Concurrent problems of economic survival obstructed the prospects for the development of firms, primainly concerned with replacing phased out products and increasing the technological level of production. The situation was aggravated by the generally unfavourable condition of legislation, taxation, and finance. The second sharp decrease in the level of innovative activity

was observed in 1995, when this indicator fell to 5.6 per cent. Since then it has decreased to 4.7 per cent in 1997.

In this way, in 1991-98, in each component of R&D provided with resources – financial, human, material and technical, information – destructive trends emerged and seriously reduced the potential of Russian science. Where would it lead? Here we itemise the most significant, in our opinion, threats to Russia's national security, caused by the disruption of its S&T potential.

First, there are global dangers: a lag in military strategy, a decrease in the level of S&T security, foreign technological dictates, an absence of allies in foreign policy, a lack of effective ways by which to achieve global competitiveness, and difficulties in the adaptation of imported S&T achievements.

Second, there are sector-specific dangers: increasing difficulties in developing the fuel and power complex, degradation of the aerospace complex and the manufacturing equipment of other rectors, stagnation in basic areas, and obstacles in the way of the conversion of defence industry.

Third, there are prospective dangers: a lag in the development of information technologies, reduced opportunities for the transition to sustainable development, and destabilisation of society in regional and social terms.

Fourth, dangers of a humanitarian and psychological nature, which are already observed in the social life of the country: the dissemination of unscientific ideas, the loss of continuity between generations, the decreasing significance of cultural values, and the aggressive behaviour of individuals and in particular of groups of people.

The question at issue is: how can we meet the challenge to the R&D sector? Obviously, it is impossible to progress on the path of isolated measures to maintain S&T potential, undertaken in response to the gravity of the situation of scientists and R&D institutions. This policy has not only failed to remove negative trends but has even strengthened the most destructive of them. The existing conditions require a goal-oriented comprehensive approach to reforms. The main task is the transition from isolated measures maintaining S&T potential to the creation of favourable conditions for the development of Russian science.

The economic crisis of 1998 dispelled many illusions about the relative prosperity of the country engendered by the exportation of fuel and raw material, the possibility of stimulating innovative processes predominantly by market self-regulation, as well as an increase in the technological level of production mainly as a result of imported developments and foreign investment. On the other hand, the capacities for survival for the Russian R&D sector based upon the potential created during the Soviet period are on the edge of

exhaustion. Therefore, a further lack of effective demand for R&D output may well lead to an irreversible degradation of most areas of research.

Under existing circumstances, the main factor in the revival of Russia's economic potential can only lie in increasing the real sector of the economy. It is necessary, without forgetting about the imperishable value of free scientific search, to pay especial attention to an evaluation of the immediate contribution that S&T can make to a quantitative and qualitative rise in production.

The answer to the question of the role of science in the restoration of Russia's economic potential is at the same time both obvious and extremely difficult. In fact, world economic science has proved that the input of scientific achievement in GDP increase is above 50 per cent (in the United States, e.g., it is 70 per cent). However, under the conditions of the most acute deficit of finance that has ever taken place in Russia, an analysis of the supposed effect from investment in R&D is very complicated.

The central place is occupied by the vital problems of transferring the economy to an innovative path of development. While declaring innovation the basis of economic development, it would be an unpardonable mistake to ignore the still existing powerful Russian scientific potential and not to use domestic developments that often are less expensive and at the same time better adapted to Russia's specific circumstances in comparison with their foreign counterparts.

The Russian R&D sector is able to offer quite a number of technologies whose further development and introduction promise large-scale and reasonably quick economic advantages.

The commercialisation of advanced Russian technologies will enable us to sufficiently increase the competitiveness of domestic goods and services. It is important to emphasise that, as a rule, contemporary scientific developments enable us to save on expensive and rare resources, including natural ones.

Under the conditions of an intensification of innovative processes, a strong influence on the sphere of production is made by choices as to the priority areas of science and technology. The priorities, once chosen, determine the technological structure of the economy for quite a long term – at least for 5-10 years. It should be remembered that the economic efficiency of scientific developments takes place not only in the sectors for which they are intended but is also multiplied through intersectoral links. Therefore, while choosing priorities, it is important to plan the development of macrotechnologies (chains of coupled technologies), which could be critical knots in achieving economic breakthrough.

It should be noted that the progressive movement of the contemporary economy is based on not only technological but also on organisational inno-

vations. In this area science seeks to have its say. The expenditure of internationally leading corporations on R&D only amounts to a value of about 5-10 per cent, and in informatics and electronics it is up to 30 per cent. Taking into account expenditure on the introduction of innovations, the share of the innovative component in expenditure is still higher. In Russia as well, it is important to ensure the integration of sector-specific R&D efforts in large business structures which are being formed. There are already examples of the establishment of financial and industrial groups associated with high-technology industries with a well-developed research and experimental base.

Among these are to be found the financial and industrial groups Formash, Avant-garde, The Russian Aviation Consortium, and High-Speed Fleet.

Within the financial and industrial group Formash, which comprises several dozens of enterprises belonging to different sectors of industry, there are 12 research institutes and design bureaux. The most important objective of this group's activities is assistance to incorporated institutions and enterprises in camying out R&D as well as introducing new equipment and technology for the manufacture of chemicals and chemical products, the manufacture of textiles and clothes, and the manufacture of machines and equipment.

The financial and industrial group Avant-garde was established in 1997 by Russia's leading research institutes and design bureaux in the field of material science (for the aerospace, nuclear, shipbuilding, electronic, and aviation industries). The aim of its creation was to help in developing the intersectoral co-operation, combination and concentration of the participants' efforts aimed at increasing the quality of S&T development of perspective materials and supplying them to the national economy.

Another important role for science in society and the economy is to expand the network of small and middle-sized innovating enterprises. The greatest success so far has been achieved by firms originating from business incubators, technology parks, technopolises, and other similar structures, whose intellectual nucleus is made up of large research institutes or higher education institutions.

Large-scale reserves of demand for science in Russian society are hidden in a greater orientation of research to regional needs. The dissemination of R&D output over the country's territory could contribute to changing the socio-economic character of the regions, by forming high-technology production clusters by concentrating the innovative potential but also by consolidating the indivisible national economic space of Russia. For these purposes the idea is to establish a network of innovative infrastructure. The creation of innovative technological centres as a basis for the national innovative system has already begun.

Basing itself on S&T, the real sector of the economy will be able not only to satisfy current innovative demand but also to help rational strategic development. In this connection it would be advisable to single out three reference points. First, the danger of losing national sovereignty and the disintegration of the country makes Russia's economic security the hot issue of the day. An application of the scientific approach will enable us to prevent the economy from further slipping into the depths and to find in the present complicated situation a narrow pathway which can lead to stabilisation.

Second, a progressive movement of Russian society and the economy is impossible without an enhancement of the informatisation processes. The role of science in the intensification of these processes is multiform. It comprises the creation of information technologies for different sectors, the development of modern information and communication systems, the generation and regulation of information arrays and streams, and mechanisms for the diffusion of S&T information.

Third, Russia can and must become an integral part of world civilisation, after adopting and implementing the internationally approved guidelines for a transition to sustainable development. In this connection, science has such tasks as the elaboration of mechanisms of rational nature management and social stabilisation taking into account national specific features, as well as the reduction of the anthropogenic burden on the environment. During the transition to a new stage of civilisation, science acts not as an aggregate of particular fields of knowledge but as an integral social institution which ensures a comprehensive and interdisciplinary approach to global, national, and particular problems.

The subject of the influence of science on the prospects for Russia's participation in international economic co-operation deserves especial discussion. Having powerful scientific potential, Russia has naturally a right to count on a considerably greater share in the world market of science-intensive products and services than the 0.3 per cent it has today. Certainly, the penetration of our developments abroad is hindered by various political, economic, and cultural obstacles, both objective and artificial. However, it is necessary to remember that, under present conditions, it is important not just to create a perfect model but also to ensure its actual marketing.

The choice of priorities as regards science and technology in many respects is correlated to the space that Russia able to occupy in international labour division in the field of high technology. The intellectual potential available in the country creates the prerequisites not only for an appreciable presence of Russia in world trade in science-intensive products but also for an active participation in more advanced forms of co-operation and integration into international research and production complexes.

The specific features of the present economic situation, as well as Russian traditions regarding the relationship between science and society, enable us to conclude that in the forthcoming years a major role in efforts to support science will as always be played by the state. For this reason, primary attention should be paid to improving the state's science and technology policies.

Among the objectives for such policies in forthcoming years, the following may be emphasised:

— The periodic correction of the priorities of S&T development, taking into account changing socio-economic demands and trends in cognition processes.

— The formation of large research and industrial complexes. They will not only serve to link science and production but also facilitate a transition from a linear to an iterative model of the R&D and innovation cycle.

— The elaboration of legal norms for compulsory deductions to be made to the special accounts of enterprises, equal to 1.0-1.5 per cent of their sales, which should be used by enterprises for R&D and innovative activities.

— The establishment of federal centres for science and high technology. This will be a way of implementing national priorities in science and technology and achieving the concentration of S&T potential in order to solve the most urgent problems.

— Incentives for investment by business entities in R&D and innovation (e.g., the possibility of introducing governmental guarantees for private investment in science-intensive areas).

— The intersectoral and interregional co-ordination of S&T and innovative activities.

The extent to which we will implement measures designed to turn science and innovation activities into a basis for economic development will in many respects determine the future of Russia for many years and decades to come.

SCIENCE AND TECHNOLOGY POLICY IN JAPAN

YUKIO SATO

INTRODUCTION

Japan is embarking upon drastic and dynamic change in her total socio-economic system at the dawn of the twenty-first century.

The Japanese Government is determined to carry out reforms in major areas. They are administrative reform, reform of the social security system, reform of financial systems, economic structural reform etc., which are necessary and inevitable, although processes of some reforms have been delayed or altered to some extent because of a severe setback to the economy.

In June 1998, the Diet passed the Administrative Reform Basic Law, which stipulates, among other things, the reinforcement of Cabinet functions and the reorganization of the current twenty-two ministries and agencies into thirteen consolidated entities starting in the year 2001. In January 1999, the government adopted the outline of bills for administrative reform in line with the Administrative Reform Basic Law. The specific bills, which are intended to be submitted to the Diet in April, will include those for establishing a Cabinet Office and other new government ministries, as well as amendments to the Cabinet Law and the National Government Organization Law.

This administrative reform, together with other reforms, is bringing about a revolutionary change in Japan which could end up by being comparable to the changes following the Meiji Restoration in 1868, and after the end of World War II in 1945. The last two revolutionary changes took place in order for Japan to catch Western countries up and to adapt to Western systems after the failures caused by isolation in Japanese social and political systems. The last two major changes were rather simple in their objectives and straightforward in their execution. Both reformations put heavy emphasis on the introduction of technologies and socio-economic systems from Western countries.

The principal driving force behind proceeding with new Japanese

reforms in the twenty-first century also arises from the necessity to adapt to the world system. However, the current situation is quite different and more complex than either of the two previous reformations. The world system surrounding Japan has itself dramatically changed and the pace of change is accelerating. Moreover, the Japanese system has been becoming interwoven with other elements in a globalizing economic system to such an extent that changes in Japan may exert great influence over world-wide developments. This is exemplified by the recent Asian economic crisis.

In addition to the change in the relationship between the Japanese and world systems, underlying conditions for the economy and society in Japan have been changing profoundly. Past practices such as lifetime employment and seniority-based wage systems are coming to an end. However, the most influential change is demographic. The birthrate in Japan has been falling dramatically in recent years. The total fertility rate dropped to 1.42 in 1995, falling below, substantially, the 2.08 threshold, the level required to maintain the current population.

Because of the trend towards fewer children, the total Japanese population is predicted to reach a peak in the year 2007 and then to decrease for the first time since World War II (fig. 1). As a result of increasing Japanese life expectancy (already the longest in the world), and a declining birthrate, the ageing of the Japanese population is proceeding at a higher rate than in any other major industrialized country (fig. 2). In addition, the proportion of the population aged 65 or over, now 15%, will become more than 30% by 2050, while the working age population (15-64 of age) which showed an annual increase every year up to 1995, is now expected to undergo a continuous decline (fig. 3).

Recent Developments in Science and Technology Policy

The government adopted a new approach to science and technology policy in the 1990s, in recognition of the important role that science and technology should play in addressing the challenges facing contemporary Japan.

In 1995, the Science and Technology Basic Law was enacted to achieve a higher standard of science and technology in order to contribute to the development of the economy and society in Japan and to the improvement of the welfare of the nation, as well as to contribute to global progress in science and technology and to the sustainable development of human society. Stating that the nation is responsible for formulating and implementing comprehensive and systematic policies with regard to the promotion of science and technology, this law laid out the scope of measures to be taken by the Government. (The outline is shown in table 1).

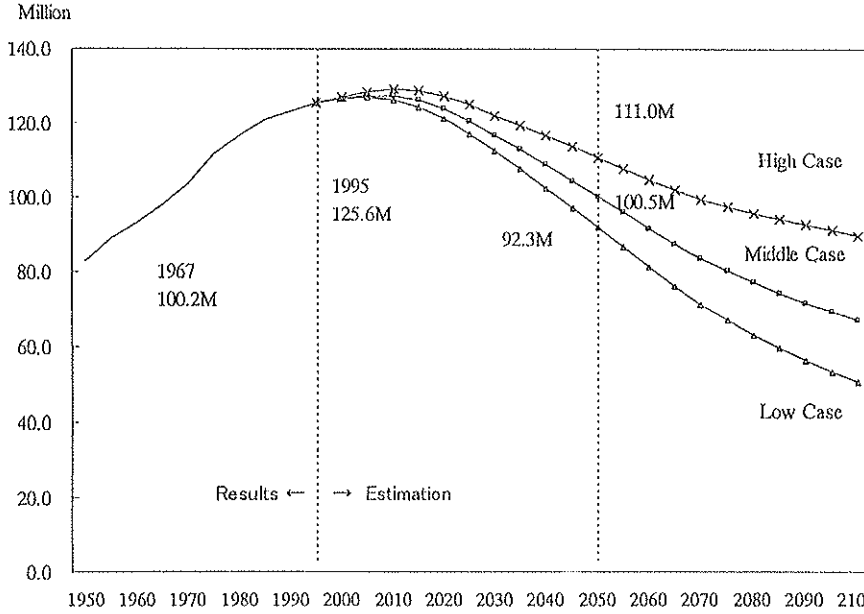
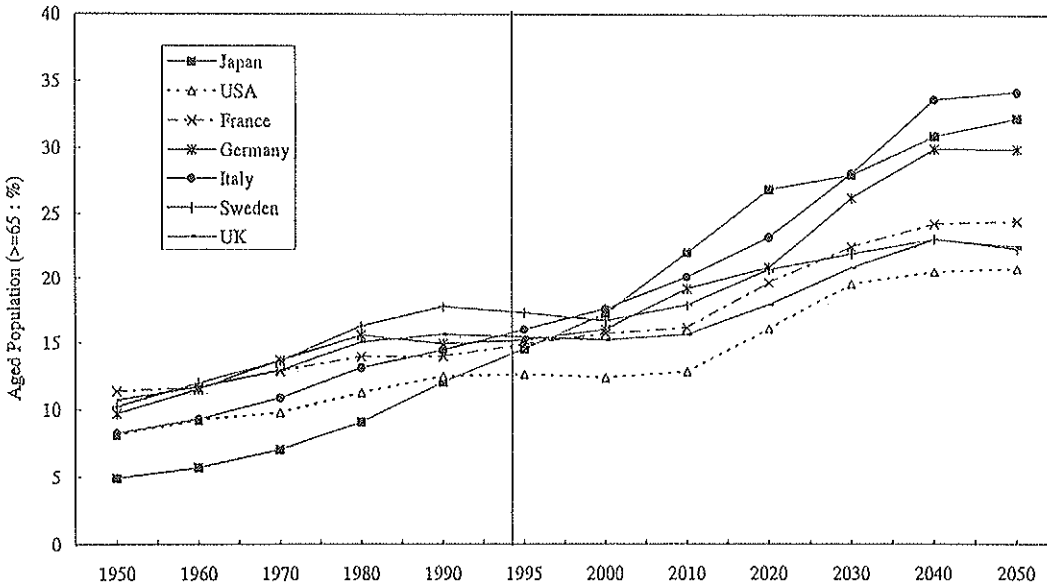


Fig. 1. Projection of the Japanese Population.



(*Source*) Japan: National Institute of Population and Social Security Research. Other: UN, World Population Prospects 1996.

Fig. 2. Projection of Ageing in Advanced Countries.

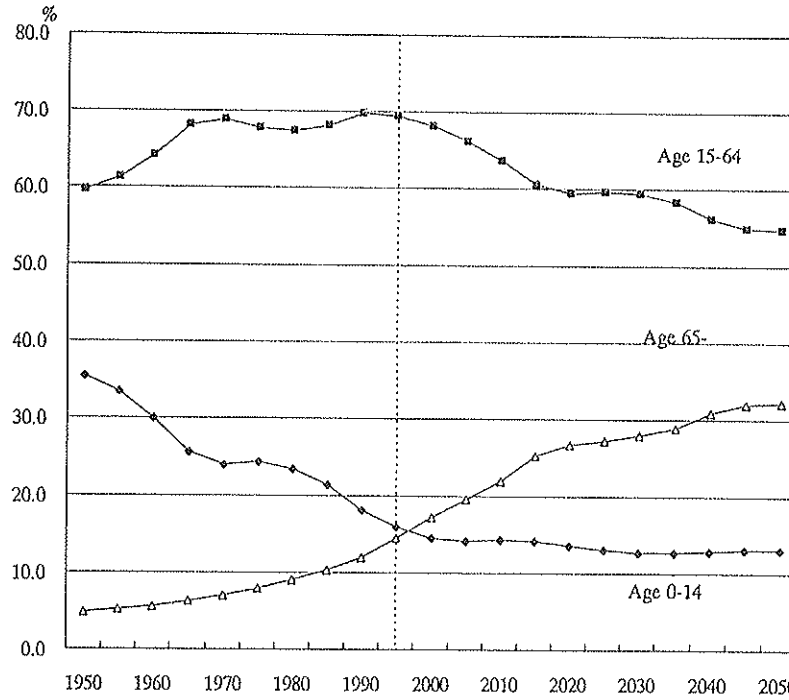


Fig. 3. Japanese Population Structure by Age Group.

In July 1996, the government decided on the Science and Technology Basic Plan based on Article 9 of the Basic Law, to implement policies for the next decade with special reference to the fiscal years 1996-2000. The plan is composed of two chapters. Chapter 1 describes comprehensive policies to promote R&D. It shows directions for R&D and improvement of systems and conditions for its promotion, including the expansion of R&D investment by the government. Chapter 2 prescribes concrete measures to be taken for the next five years based on the policies laid down in Chapter 1 (the outline is shown in table 2).

The Administrative Reform Basic Law stipulates, among other things, that the Science and Technology Agency and the Ministry of Education, Science, Sports and Culture (Monbusho) are to be merged into one new Ministry of Education, Science and Technology. The law also establishes a new General Council for Science and Technology to replace the existing Council for Science and Technology (CST). The secretariat of the new General Council for Science and Technology (GCST) will be based in the Cab-

Table 1. *The Outline of the Science and Technology Basic Law.*

(1) *Guidelines for S&T Promotion:* — Development of Researchers' Creativity; — Harmonious Development of Basic, Applied and Developmental Research; — Promotion of S&T in harmony with Human Life, Society and Nature.
(2) *Responsibilities of National and Local Government in Promotion S&T.*
(3) *Formulation of S&T Basic Plan by the Government through consultation with the Council for S&T, in order to Promote S&T Policies comprehensively and systematically:* — Taking Necessary Measures to Secure Funds for Implementation of the Basic Plan.
(4) *Measures taken by the National Government:* — Balanced Promotion of diversified R&D; — Securing and Training Researchers and Technicians; — Improvement of Research Facilities; — Promotion of Information Intensive Research; — Promotion of R&D Exchanges.

Table 2. *The Outline of the Science and Technology Basic Plan.*

(1) *The Basic Aims of R&D:* — Intensive promotion of R&D applicable to socio-economic needs such as creation of new industries, solving global problems, construction of a comfortable community; — Active promotion of basic research, aiming at the discovery of new laws and principles, the creation of original theories, and the prediction and discovery of unknown phenomena.
(2) *Constructing a New R&D System:* — Constructing a new R&D system to facilitate creative R&D activities; — Constructing a cooperation and exchange system beyond sectors, regions and nations; — Impartial evaluations.
(3) *Expansion of R&D investment by the Government:* — Expected S&T budget total for 5 years from 1995 to 2000 is about 17 Trillion yen. (about 130 billion US$).

inet Office. Both changes will have the effect of strengthening science and technology policy coordination. The GCST will also assume responsibility for the humanities and the social sciences (fig. 4). Moreover, it is expected that more than half of Japan's national research institutes, now attached to ministries and agencies, will become "Independent Administrative Corporations" which will have greater independence from the government.

A quick review of the historical evolution of the structure for S&T in Japan shows the drastic nature of these recent developments. This is exemplified by the Basic Law to promote science and technology, the first of its kind to be enacted in Japan's history. The change in the Japanese administrative structure would alter two main principles concerning S&T administration which have existed for more than forty years.

Firstly, the administrative reform would put an end to a unique system in which an agency headed by a Cabinet Minister, the Science and Technology Agency (STA), has been dedicated to formulating comprehensive S&T policies and to coordinating S&T policies across ministerial boundaries. Here it should be noted that S&T policies in Japan are implemented by various ministries and agencies according to their respective missions.

Secondly, inclusion of the humanities and the social sciences in "science and technology" would bring about new perspectives and complexities in the formulation and implementation of S&T policies. The compound word "science and technology" in Japanese, together with the exclusion of those disciplines relevant only to the humanities and the social sciences, have produced a somewhat unique notion which has characterized Japan's remarkable development in science and technology since the end of the Second World War. The need for increased harmony between science and technology and human society requires more involvement of the humanities and social sciences in the formulation of S&T policy.

In addition to these changes, a new system for nation-wide science and technology research evaluation, which has already started, would exert profound influence on the research community and researchers of the national institutes and universities.

The Science and Technology Basic Plan specified the establishment of a national guideline to implement evaluation for governmental R&D. The National Guideline on the Method of Evaluation for Government R&D, drafted by the Council for Science and Technology (CST) and approved by the Prime Minister in August 1997, applies to all government-funded R&D. It therefore applies to R&D implemented by national research institutions, national universities and special public corporations. It also applies to R&D implemented by private sector institutions acting as partners in joint research with government funding, to R&D implemented by institutions of

Proposed Administrative Reform
by the Administrative Reform Basic Law

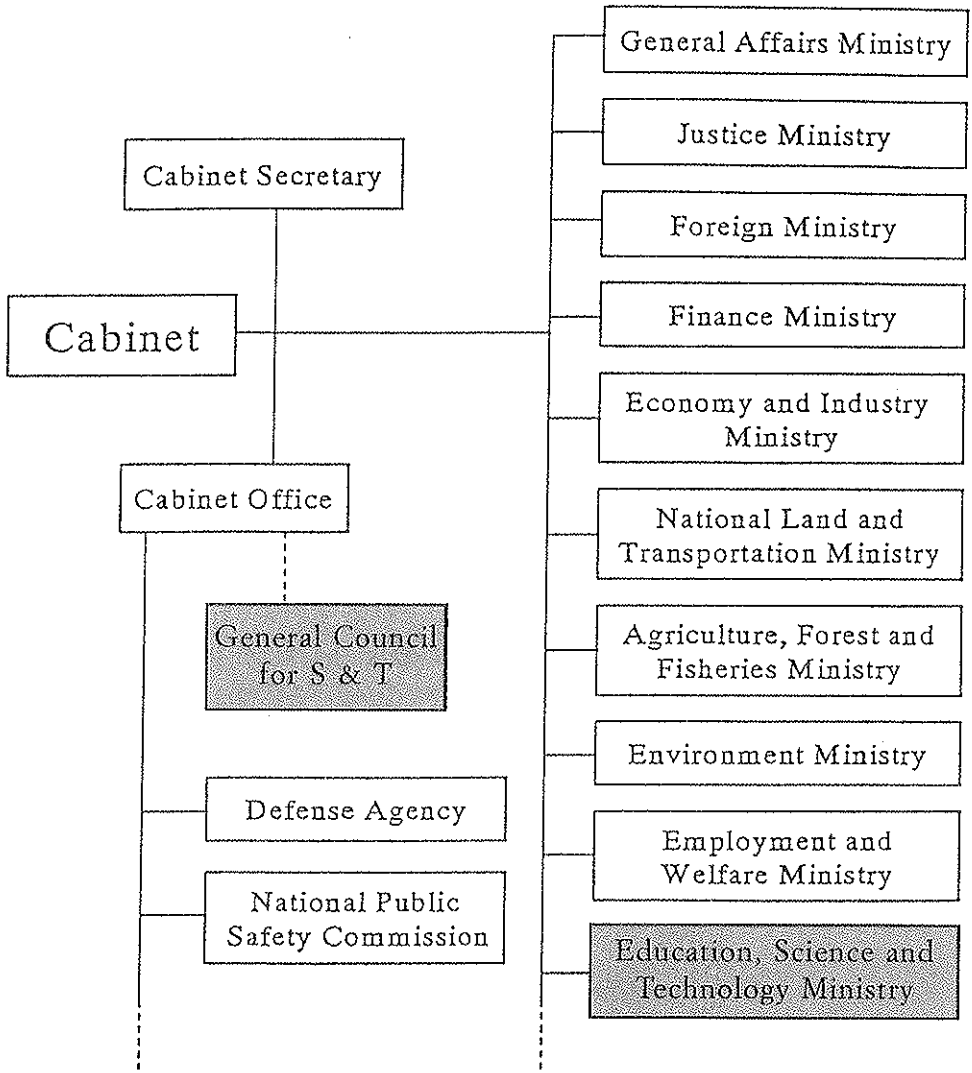


Fig. 4. Administrative Reform.

local governments with support from government funding, as well as to R&D implemented overseas funded by the Japanese government.

The introduction of fair and strict evaluation into the governmental R&D system responds to two major growing requirements. One is to promote R&D in response to social and economic needs under severe financial conditions, and the other is to gain the public's deep and broad understanding of the need for the promotion of science and technology. A new system for S&T policy in line with the legal reforms described thus far would induce more explicit distinctions between different functions involved in the formulation and implementation of S&T policies.

Each ministry or agency responsible for policy formulation has national research institutes whose functions are to implement R&D in the field of its jurisdiction. These research institutes together with R&D related public sector corporations, although their major role is the implementation of R&D, have greatly contributed to the formulation of national S&T policies through the exchange of information and personnel on a regular and/or *ad hoc* basis. Transformation of national research institutes into independent entities would change the basis of the existing relationship between policy formulation and implementation organizations to a more contractual one. Policies and missions set by ministries would be carried out by the research institutes through R&D contracts.

Profile of Japanese R&D Activities

Japan's total expenditure on R&D during the fiscal year (FY) 1997 stood at 15,746 billion yen (14,492 billion yen for the natural sciences), an increase of 4.4% (6.1% for natural science) over the preceding year (table 3). This is a consecutive three-year increase. The total R&D expenditures in FY1993 and FY1994 declined due to lower investment by the private sector. This was due to an economic slowdown, following a long period in which Japan had enjoyed steady growth in the level and scope of national science and technology activities.

Of the total R&D expenditures (for the natural sciences) in FY1997, 10,658 billion yen was spent by companies, an increase of 6.0%; 1,906 billion yen was spent by universities and colleges, up by 1.2%; and 1,927 billion yen by research institutions, up by 1.5% compared with FY1996.

The ratio of total R&D spending to GDP in FY1997 was 3.12%, up by 0.12 point from FY1996. The ratio fell for four consecutive years in succession from FY1991, partly due to declining private R&D expenditure, but then jumped up in FY1996. Japan continues to maintain the highest level of R&D spending among major industrized countries (fig 5).

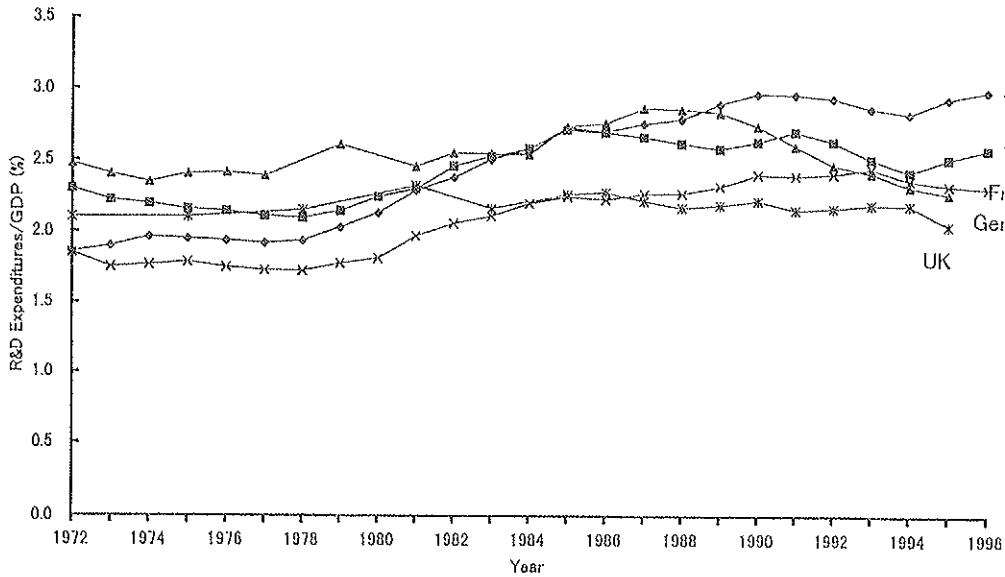
Table 3. *Outline of the National Guideline on the Method of Evaluation for Government R&D.*

(1) *Purpose of Evaluation:*
— Effective implementation of R&D evaluation;
— An efficient allocation of limited government R&D funds;
— Realization of open/flexible/competitive R&D environment;
— Seeking public support for R&D activities.
(2) *Approach to Evaluation:*
— Clarification of evaluation criteria and processes;
— Introduction of external evaluation;
— Disclosure of relevant information on evaluations;
— Appropriate utilization of evaluation results.
(3) *Matters Demanding Care:*
— Avoidance of excess burden which accompanies evaluation;
— Appropriate consideration of character of R&D;
— Utilization of numerical indices;
— Evaluation of technical examination and R&D which have difficulty producing results in the short-term;
— Harmonization of R&D with human lifestyle/society and nature.
(4) *Evaluation of R&D Institutions (Institution Evaluation):*
— National Research Institutions: evaluation result must be reflected in improvement of management of the institution;
— Universities: Establish further self-examination/evaluations.

The private sector spent 12,494 billion yen in FY1997, accounting for 79.4% of the total, while the public sector including local governments provided 3,204 billion yen, accounting for 20.4% of the total (fig 6). Shares by type of activity were 14.3% for basic research, 24.4% for applied research, and 61.3% for development research. The structure of R&D expenditure in industry, government, companies, research institutions, and universities and colleges is distinctly different. Industry spends more on development due to the nature of corporate activities and the relatively high costs of commercialization. Universities and colleges give priorities to basic and applied research Research institutions fall between the two categories (fig. 7).

Thus the low proportion of R&D expenditure devoted to basic research is to be attributed to the low share held by government R&D expenditure.

In Japan, science and technology policies are implemented by various ministries and agencies according to their respective missions. Therefore, almost all ministries and agencies have their own research institutions and budgets to do R&D works related to their missions (fig. 8).

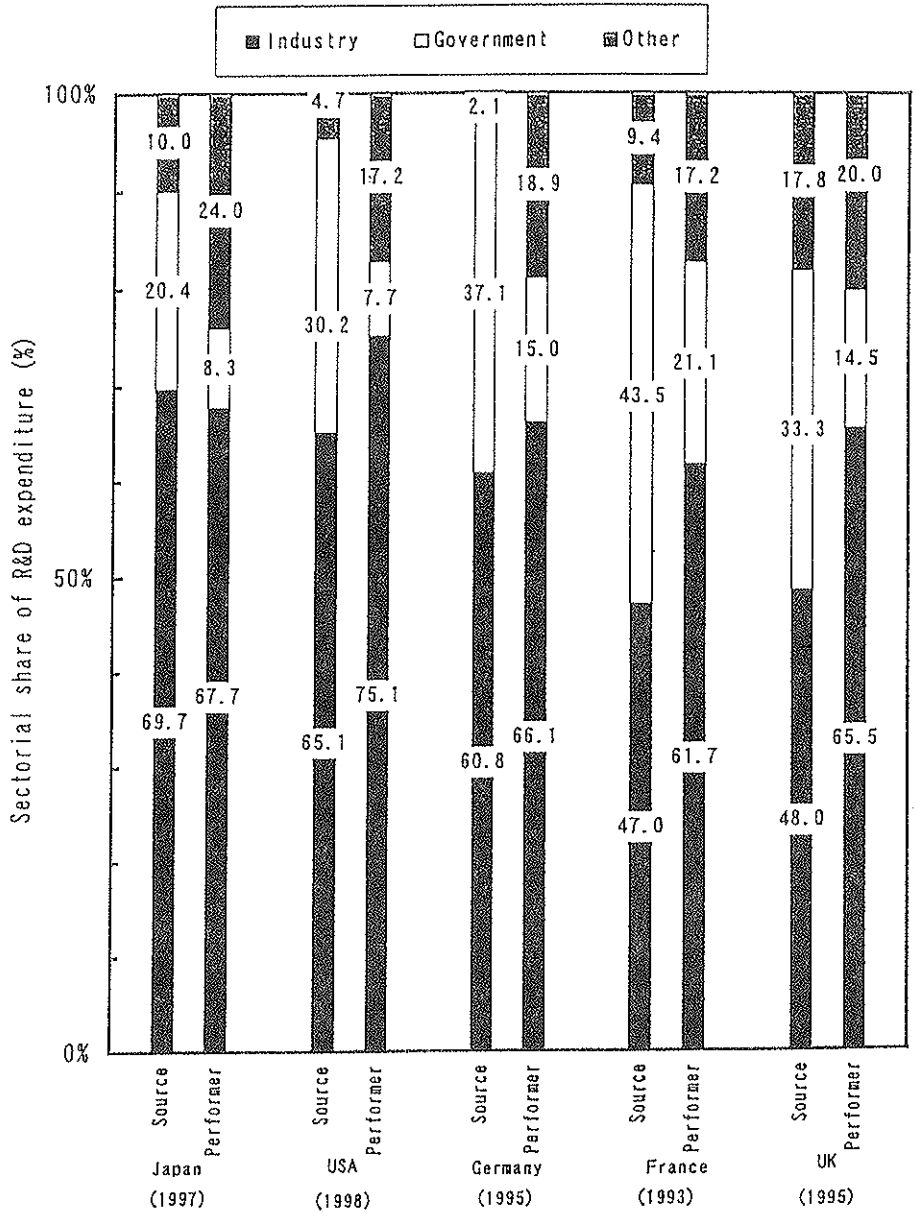


Sources: Japan – Management & Coordination Agency, *Report on the Survey on S&T Research.*
USA – National Science foundation, *National Patterns of R&D Resources: 1998.*
Germany – Bundesministerium fur Bildung und Forschung, *Bundesbericht Forschung: 1996.*
France – OECD, *Basic Science and Technology Statistics 1996.*
UK – *Forward Look 1996.*

Fig. 5. Trends in Ratio of R&D Expenditure to GDP in Selected Countries.

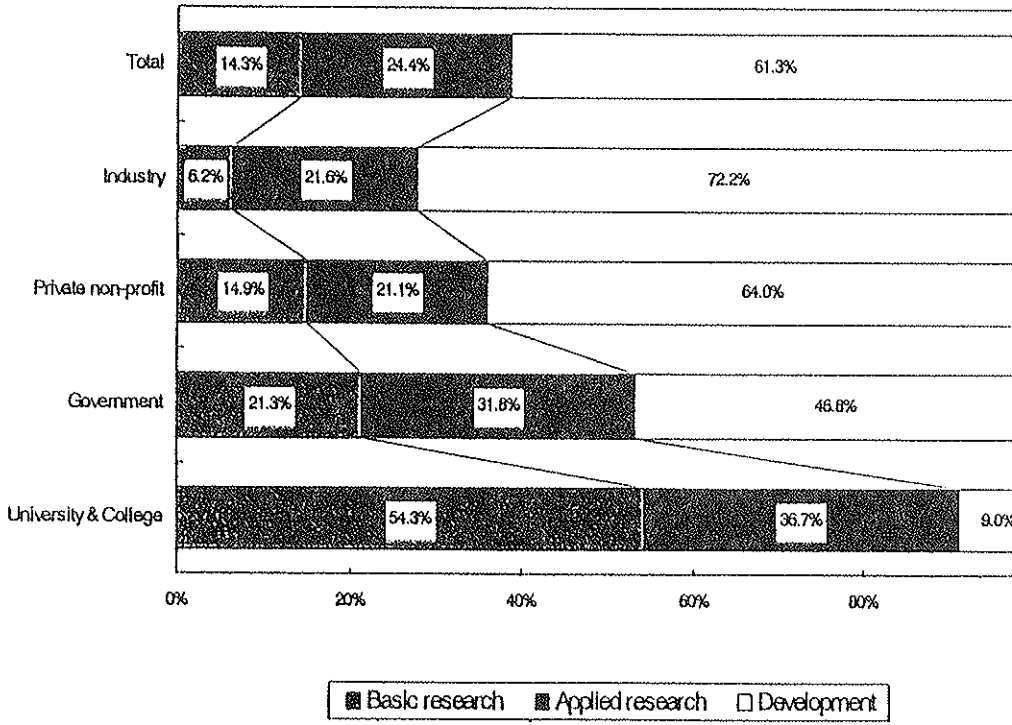
The description of the government science and technology budget is presented in fig. 9. In FY1999, the total science and technology budget of the govermnent is 3,151 billion yen. Of that amount, 42.5% is designated for Monbusho, most of which finances national universities. It also includes grants to university researchers as well as grants to private universities. 24.3% of the total science and technology budget is allocated to the Science and Technology Agency (STA). The STA carries out R&D in nuclear energy and space development and promotes basic research in important areas. 16% of the total is for the Ministry of International Trade and Industry (MITI). The Defense Agency of Japan accounts for 4.6% of the total S&T budget.

The Science and Technology Basic Plan estimates the total S&T budget for the 5 years from 1996 to 2000 to amount to 17 trillion yen. Following this ambitious plan, the national budget for science and technology has increased significantly for four consecutive years, reaching a cumulative total of making the sum of these budgets in 4 years 78.1% of the overall planned total (fig. 10).



Sources: Japan – Management & Coordination Agency, *Report on the Survey on S&T Research.*
USA – National Science foundation, *National Patterns of R&D Resources: 1998.*
Germany – Bundesministerium fur Bildung und Forschung, *Bundesbericht Forschung: 1996.*
France – OECD, *Basic Science and Technology Statistics 1996.*
UK – *Forward Look 1996.*

Fig. 6. Sectorial Share of R&D Financing and Expenditure by Source and Performer in Selected Countries.



Sources: Management & Coordination Agency, *Report on the Survey on S&T Research.*

Fig. 7. Percentage Distribution of R&D Expenditure by Character of Activity in Each Sector (FY1997).

As for the number of researchers, countries use different definitions and statistical measures. Therefore, simple comparisons may not be adequate, although it is useful to grasp general trends of human resources in research. The number of researchers in Japan as of April 1st, 1998 was 614,000 (705,000 including researchers in the social sciences and humanities), showing an increase of 1.2% (1.3% including researchers in social science and humanities) from the preceding year (fig 11).

The number of researchers by sector in 1997 is estimated as follows:

— Industry: 400,000 (65.9%);
— Universities and colleges: 164,000 (27.0%);
— Government research institutions (including public-sector corporations): 28,000 (4.7%).

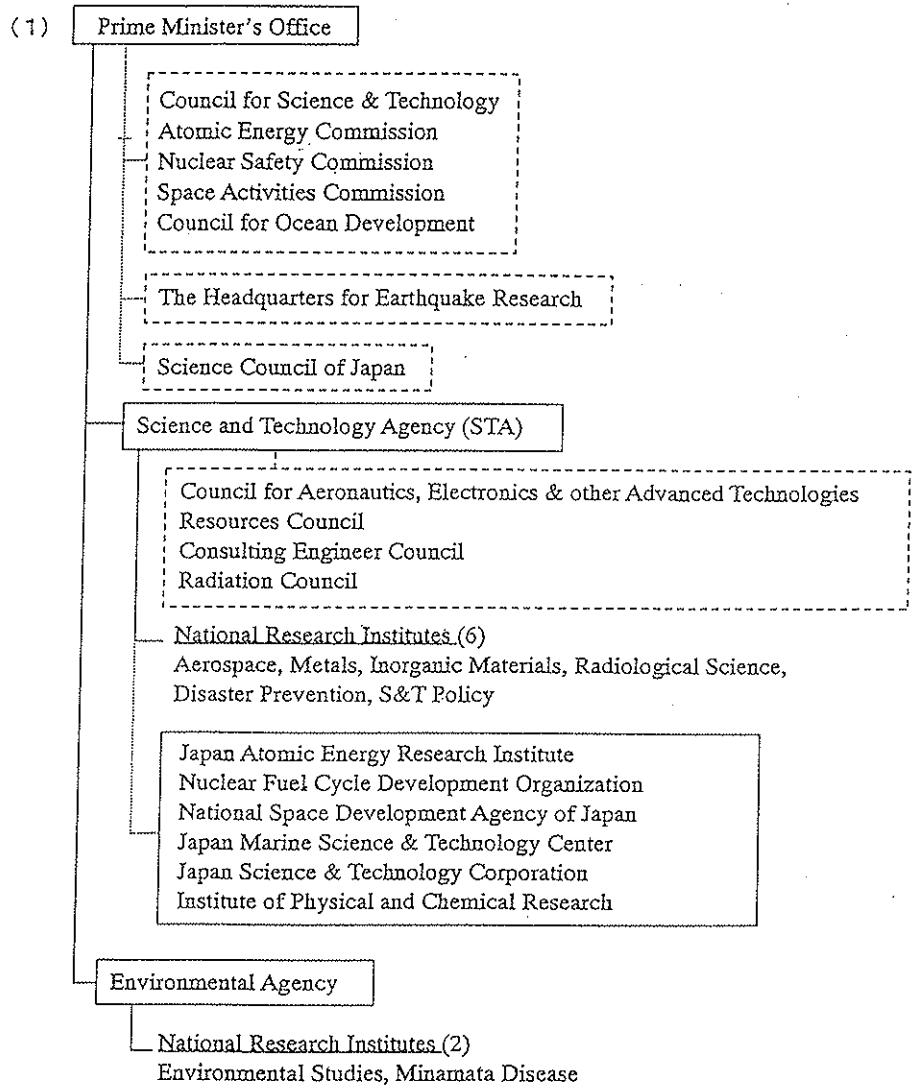


Fig. 8 (1). Major Administrative Organizations for Science and Technology.

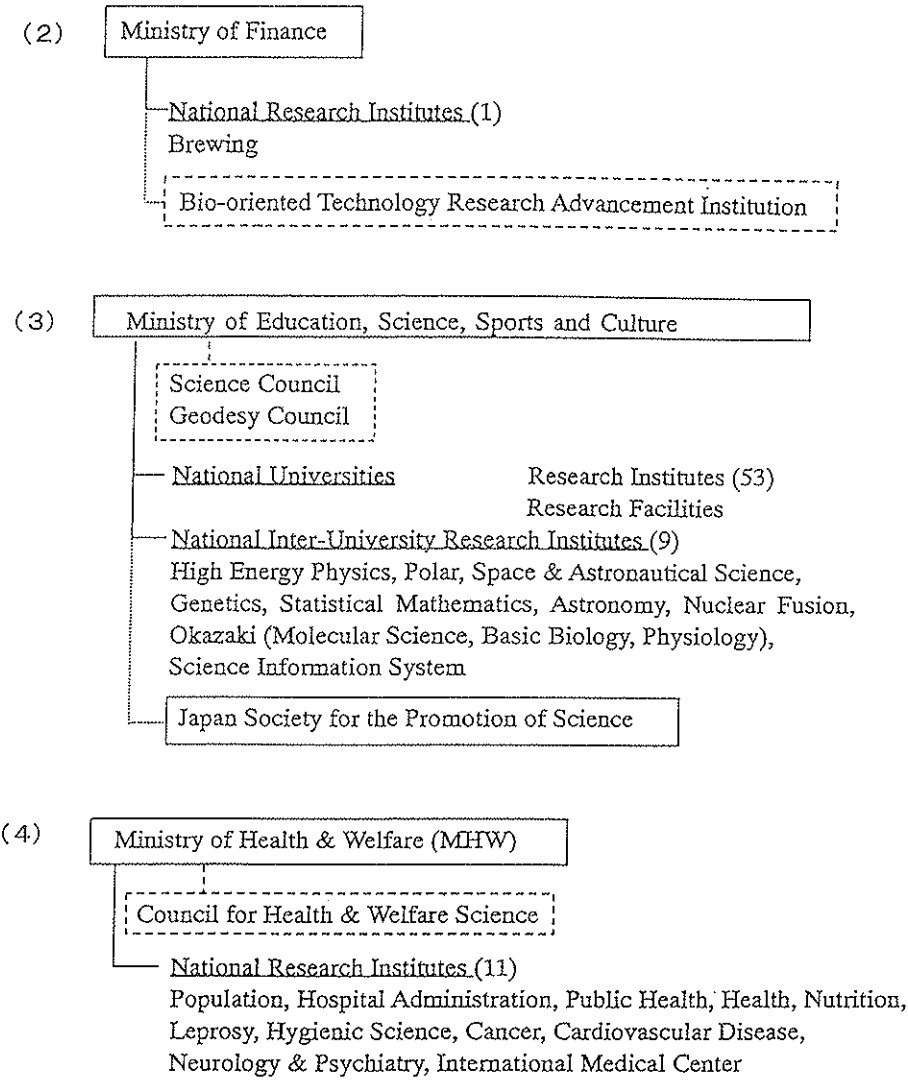

(2)
Ministry of Finance
National Research Institutes (1)
Brewing
Bio-oriented Technology Research Advancement Institution
(3)
Ministry of Education, Science, Sports and Culture
Science Council
Geodesy Council
National Universities
Research Institutes (53)
Research Facilities
National Inter-University Research Institutes (9)
High Energy Physics, Polar, Space & Astronautical Science,
Genetics, Statistical Mathematics, Astronomy, Nuclear Fusion,
Okazaki (Molecular Science, Basic Biology, Physiology),
Science Information System
Japan Society for the Promotion of Science
(4)
Ministry of Health & Welfare (MHW)
Council for Health & Welfare Science
National Research Institutes (11)
Population, Hospital Administration, Public Health, Health, Nutrition,
Leprosy, Hygienic Science, Cancer, Cardiovascular Disease,
Neurology & Psychiatry, International Medical Center

Fig. 8 (2, 3, 4).

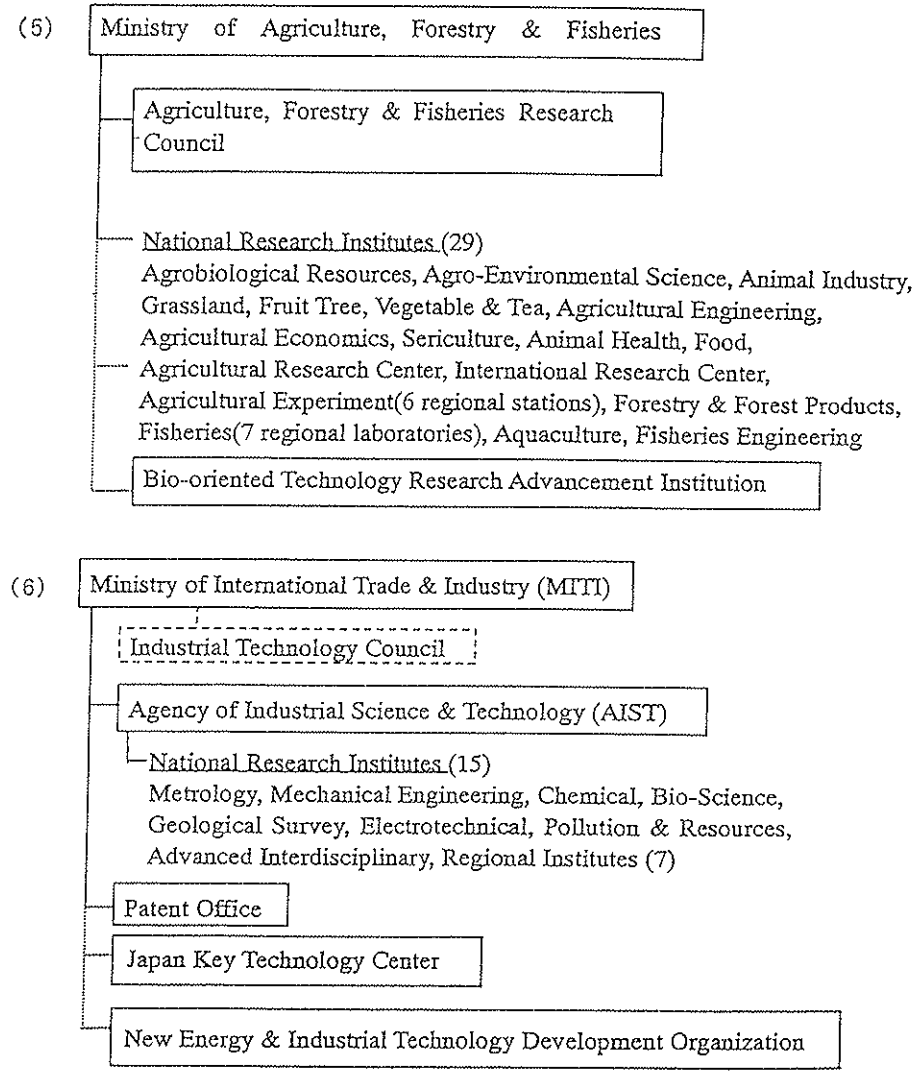

(5)
Ministry of Agriculture, Forestry & Fisheries
Agriculture, Forestry & Fisheries Research Council
National Research Institutes (29)
Agrobiological Resources, Agro-Environmental Science, Animal Industry, Grassland, Fruit Tree, Vegetable & Tea, Agricultural Engineering, Agricultural Economics, Sericulture, Animal Health, Food, Agricultural Research Center, International Research Center, Agricultural Experiment(6 regional stations), Forestry & Forest Products, Fisheries(7 regional laboratories), Aquaculture, Fisheries Engineering
Bio-oriented Technology Research Advancement Institution
(6)
Ministry of International Trade & Industry (MITI)
Industrial Technology Council
Agency of Industrial Science & Technology (AIST)
National Research Institutes (15)
Metrology, Mechanical Engineering, Chemical, Bio-Science, Geological Survey, Electrotechnical, Pollution & Resources, Advanced Interdisciplinary, Regional Institutes (7)
Patent Office
Japan Key Technology Center
New Energy & Industrial Technology Development Organization

Fig. 8 (5, 6).

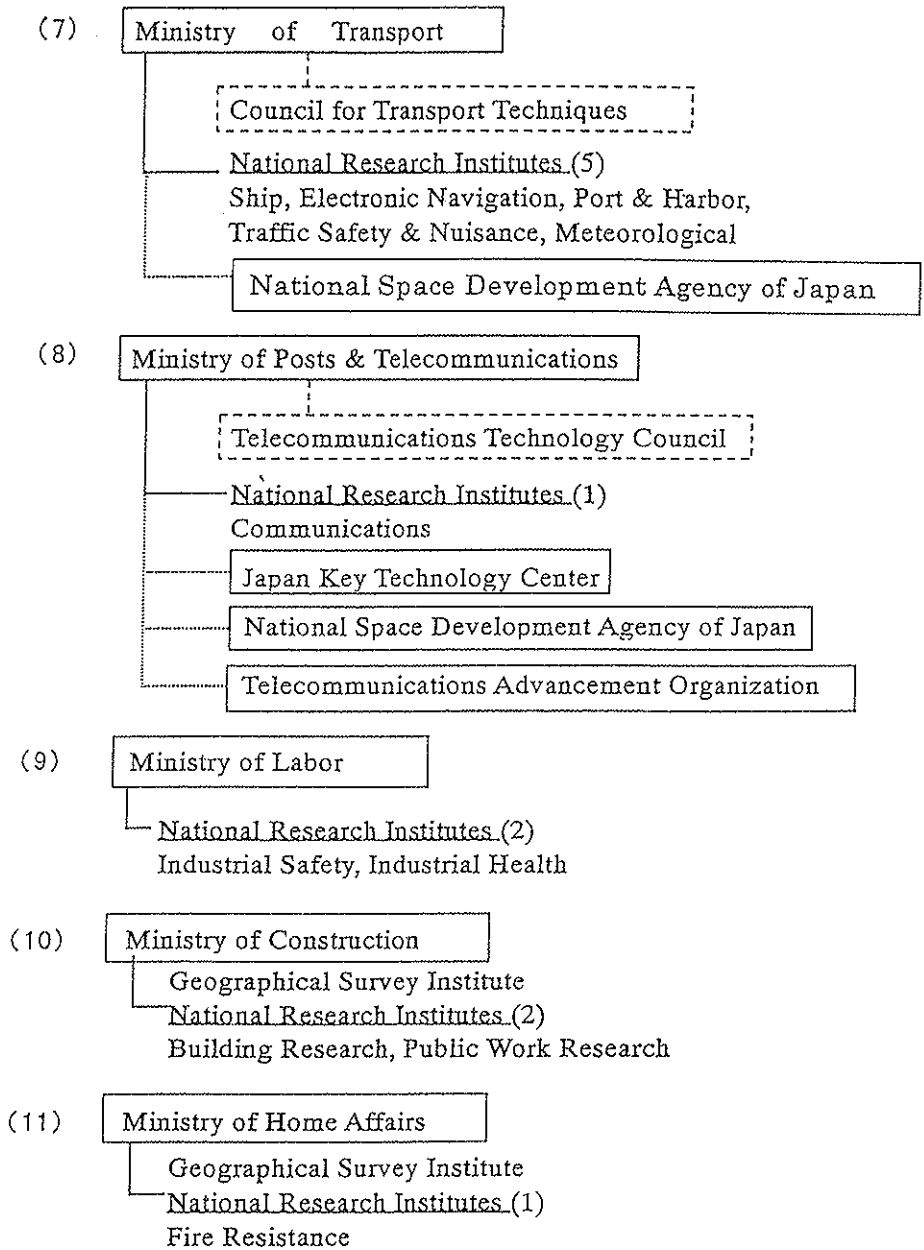
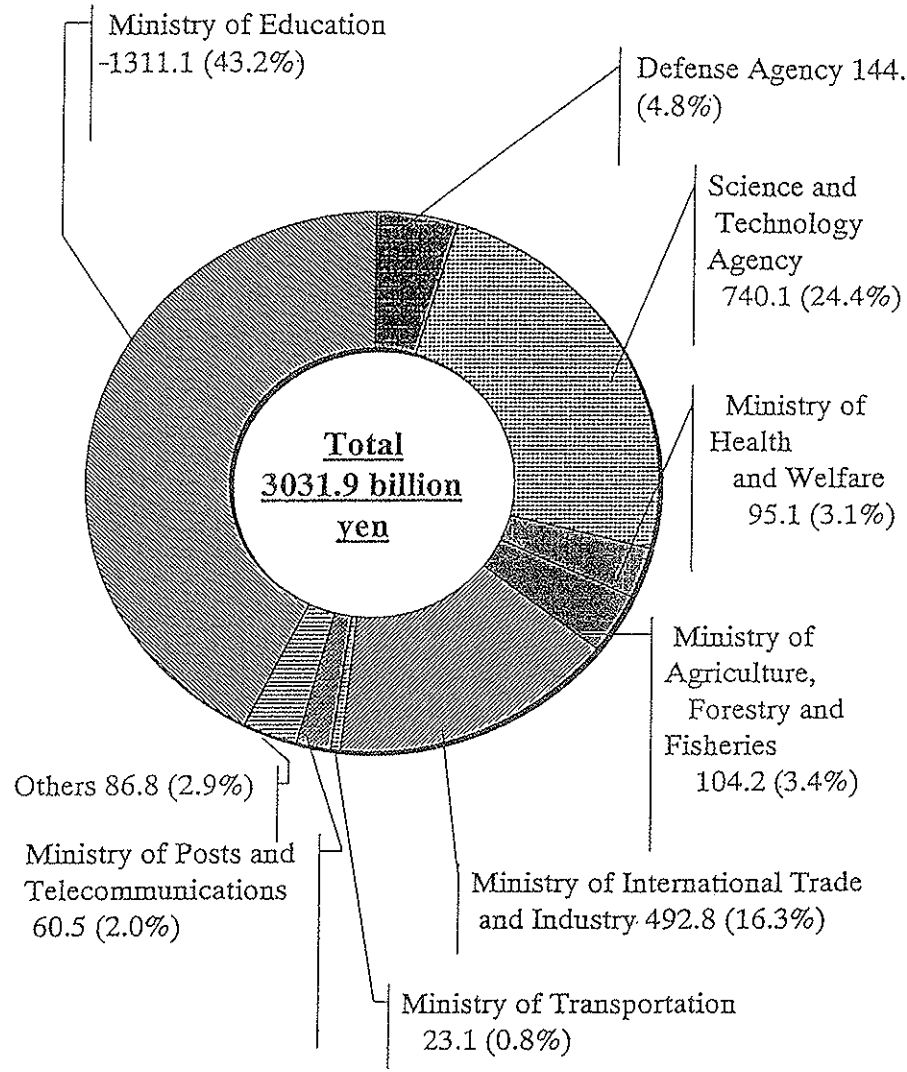


Fig. 8 (7, 8, 9, 10, 11).



Unit: Billion yen: the figures in parentheses show the percentages of the total amount.
Note: 26.0 billion yen, appropriated for the Japan Key Technology Center, is included in the budgets of the Ministry of International Trade and Industry and the Ministry of Posts and Telecommunications, in duplication (Duplications are eliminated in totaling).

Fig. 9. FY 1998 Budget for Science and Technology by Ministries and Agencies.

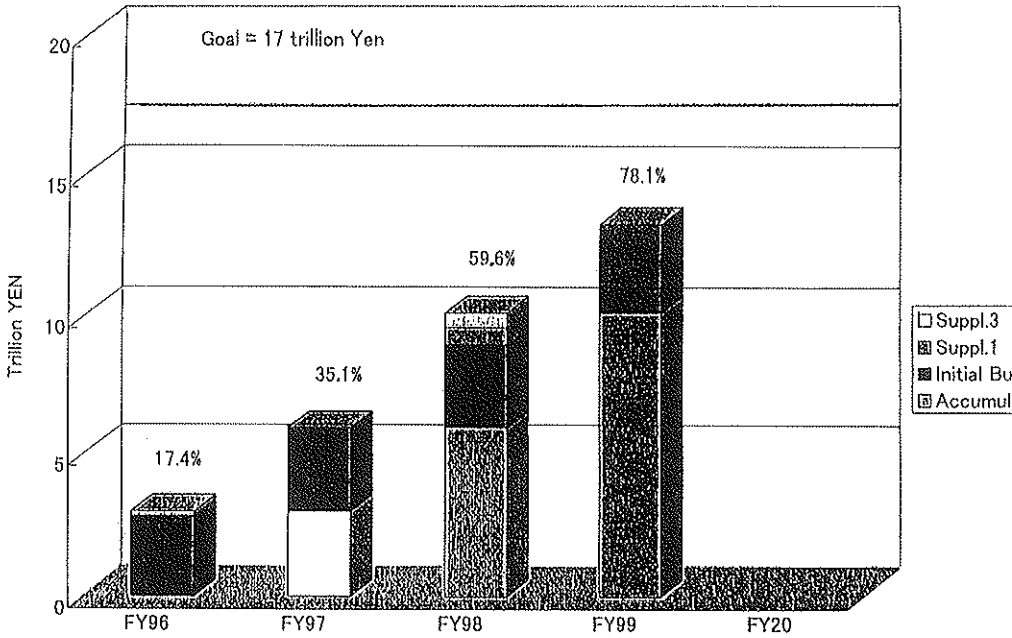


Fig. 10. Science and Technology Expenditure during the Science and Technology Basic Plan Period (5 years from FY 1996).

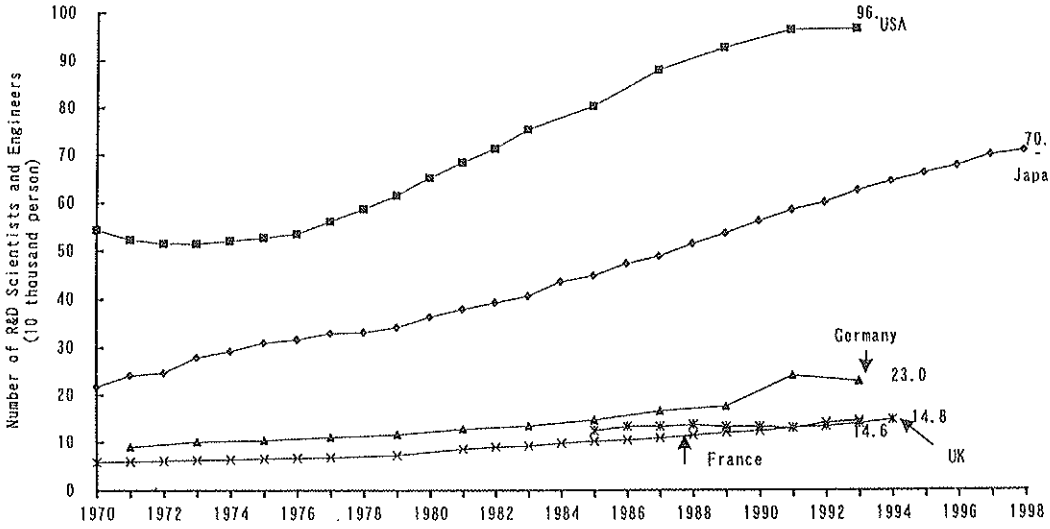
The Council for Science and Technology and the Evolution of Comprehensive S&T Policy

The 1995 Basic Science and Technology Law states that, in formulating the Basic Science and Technology Plan, the government must go through the Council for Science and Technology (CST) beforehand. The main task of the Council for Science and Technology (CST) is to advise the Prime Minister on the formulation of:

— general science and technology policy,
— comprehensive long-term research goals, and
— basic policy measures for promoting research in important areas.

Under the law establishing the Council for Science and Technology, the Prime Minister must consult the CST in cases where inter-ministerial coordination is necessary. If there is a report in response to that consultation, its recommendation must be taken into account.

The CST consists of the Prime Minister, as chairperson, four Cabinet ministers (from the Ministry of Finance, Monbusho, the Economic Plan-



Sources: Japan – Management & Coordination Agency, *Report on the Survey on Science and Technology Research.*
USA – National Science foundation, *National Patterns of R&D Resources: 1998.*
Germany – Bundesministerium fur Bildung und Forschung, *Bundesbericht Forschung: 1996.*
France – OECD, *Basic Science and Technology Statistics 1996.*
UK – *Forward Look 1996.*

Fig. 11. Trends of Number of R&D Scientists and Engineers in Selected Countries.

ning Agency, and the Science and Technology Agency (STA)), the chairperson of the Science Council of Japan and five distinguished leaders from the academic and industrial world.

These five members are appointed by the Prime Minister with the consent of the Diet. Because it is composed in this way, the S&T policies recommended by the CST have been well implemented by various ministries and agencies as if they were Cabinet decisions. In practice, the CST is a quasi-advisory, quasi-executive body.

Plenary meetings of the CST are usually held twice a year with the Prime Minister in the chair. Cabinet ministers other than those listed above may also attend plenary meetings. A panel of the CST, the Committee on Policy Matters deliberates on important policy issues and meets regularly twice a month. The Committee is composed of the members of the CST (except Cabinet ministers) and of several distinguished experts from academic, industrial and administrative spheres appointed by the Prime Minister. In addition to the Committee on Policy Matters, there are several panels composed of selected non-ministerial members of CST plus additional co-opted experts in various fields to consider important policy issues.

In designing a new framework for S&T policy, it is necessary and instructive to review the old framework in a positive way. Since a review of the evolution of the whole Japanese government structure in S&T policy would require several volumes, only a brief overview is given here. The focus here is on the framework for the formulation and coordination of comprehensive policies and the role of the CST.

It was in the period from the mid 1950s to the mid 1960s that the main organizational framework for the S&T policy-making system in Japan was established. This reflected a growing recognition of the importance of developing indigenous technological capacity rather than depending on technology transfer from advanced countries.

In 1956, the Science and Technology Agency (STA) was established to promote national science and technology. The responsibilities of the STA were defined as follows:

— to formulate comprehensive S&T policy;
— to coordinate the S&T activities of various ministries and agencies;
— to promote the development of comprehensive R&D projects to fulfil national policies on nuclear energy, aerospace, material science and processing, and other areas as required.

However, there remained a perceived need to integrate research at universities with overall national S&T policy, since academic research fell outside the jurisdiction of the STA. In order to address this need, the CST was established in 1959 as the supreme advisory body to the Prime Minister encompassing both academic and government research policy. In the following year the CST presented a recommendation entitled "Comprehensive and Fundamental Measures for the Development of Science and Technology for the Next Decade". The recommendation proposed a policy to increase human resources available for science and engineering and to intensify R&D efforts in order to catch up with advanced Western countries. This recommendation was the first overall policy framework for science and technology in Japan.

Changes of emphasis in Japanese Science and Technology policy thereafter have been brought about by the main recommendations of the CST. In 1971, the GST proposed measures in response to problems encountered in the period of high economic growth such as environmental pollution. The recommendation made in 1977 stressed the need to strengthen responsiveness to sudden changes in the international environment such as the oil crisis. It also stressed the need to develop government policies which promoted concern for a better quality of life.

In 1986, the Cabinet approved the General Guideline for Science and Technology Policy based on a recommendation made by the CST, recogniz-

ing that Japan had become one of the technologically advanced countries. This policy was based on three major objectives: the promotion of creative science and technology, the promotion of science and technology in harmony with society, and an emphasis on the international aspects of science and technology. An updated Guideline was decided by the Cabinet in 1992, based on the eighteenth recommendation of the CST. The new guideline sets the following goals: coexistence of humans in harmony with the earth, the expansion of intellectual stock, and the construction of a changing society where people can live with peace of mind.

In accordance with this guideline and the Science and Technology Basic law, based on the recommendation of the CST, the government established the Science and Technology Basic Plan. In addition to these recommenda-

Table 4. *R&D Expenditure.*

R&D Expenditure by Sector (Natural Science Only) (Unit: bil. yen)

Sector \ FY	1990	1991	1992	1993	1994	1995	1996	1997
Industry	9,267	9,743	9,561	9,054	8,980	9,396	9,881	10,658
Research Institutes	1,416	1,516	1,661	1,789	1,753	1,921	1,899	1,927
Universities	1,406	1,461	1,566	1,686	1,685	1,875	1,883	1,906
Total	12,090	12,720	12,788	12,528	12,419	13,191	13,664	14,492

Government Funding (Natural Science Only) (Unit: bil. yen)

FY	1990	1991	1992	1993	1994	1995	1996	1997
Government	1,990	2,137	2,306	2,558	2,504	2,866	2,723	2,776
Government (%)	16.5	16.8	18.0	20.4	20.2	21.7	19.9	19.2
(Exclusive of defense R&D)	(15.6)	(15.9)	(17.0)	(19.3)	(19.0)	(20.6)	(18.7)	(18.2)

Note: Because of rounding under 1 billion yen, cumulative amounts and total amounts may not be identical.

Table 5-1. *Recommendations by the Council for Science and Technology.*

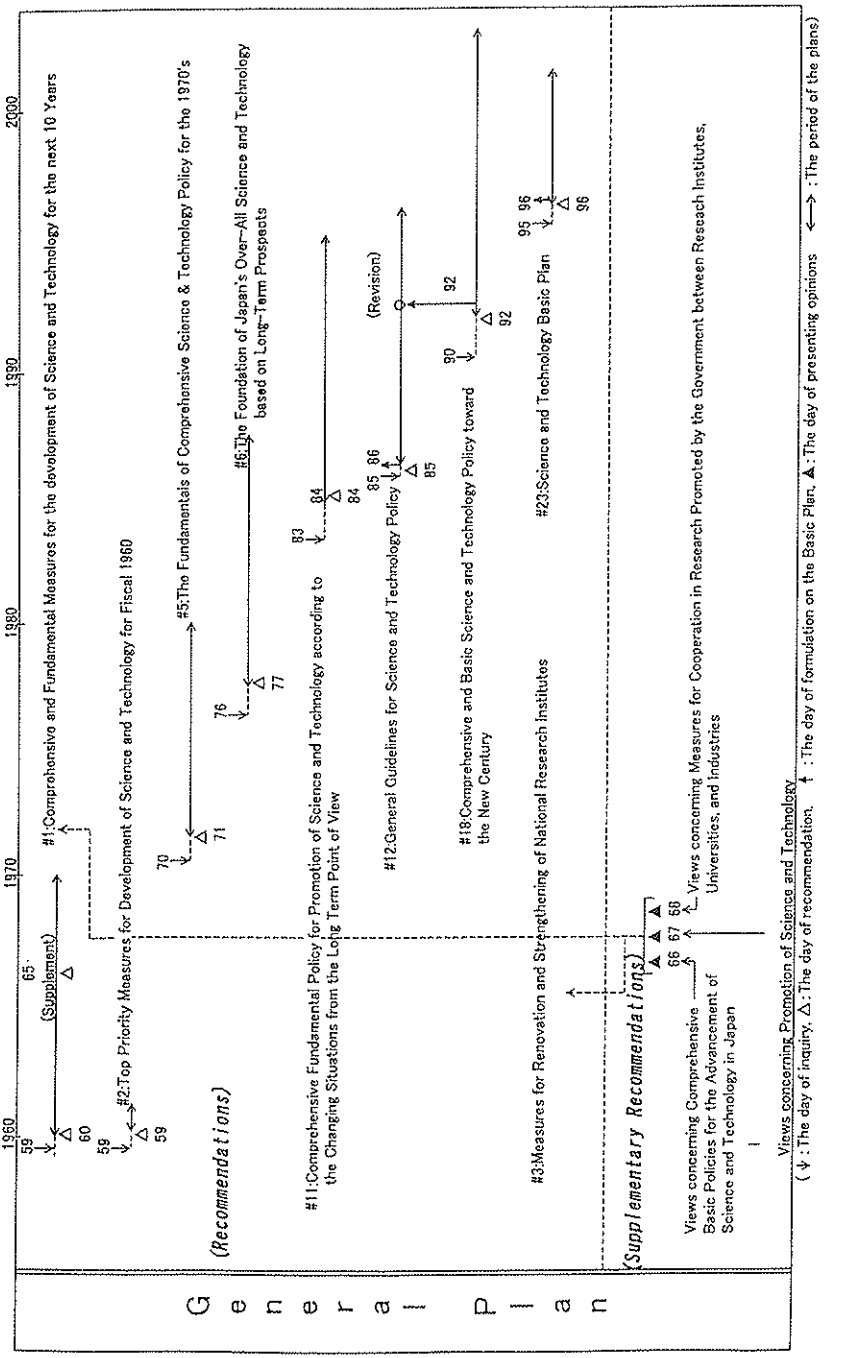


Table 5-2. *Recommendations by the Council for Science and Technology.*

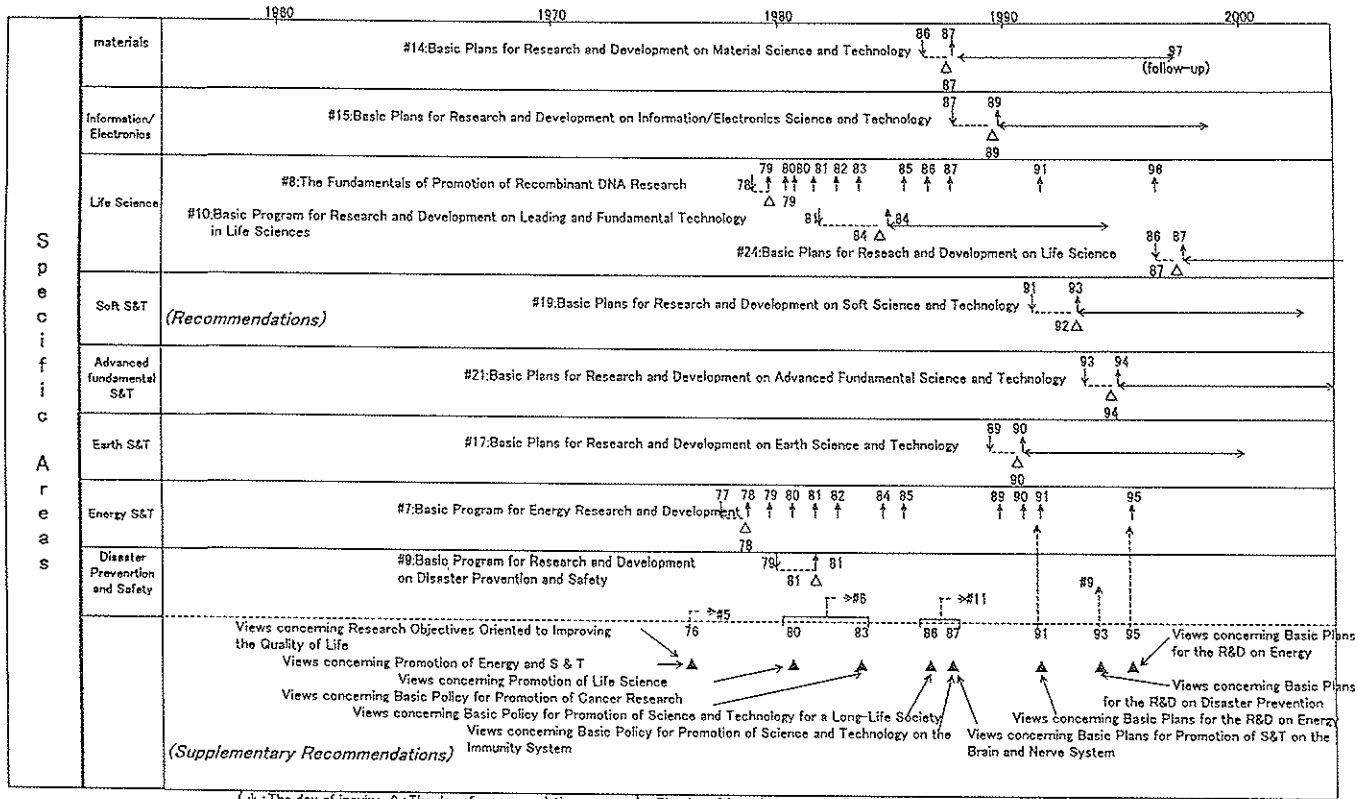


Table 5-3. *Recommendations by the Council for Science and Technology.*

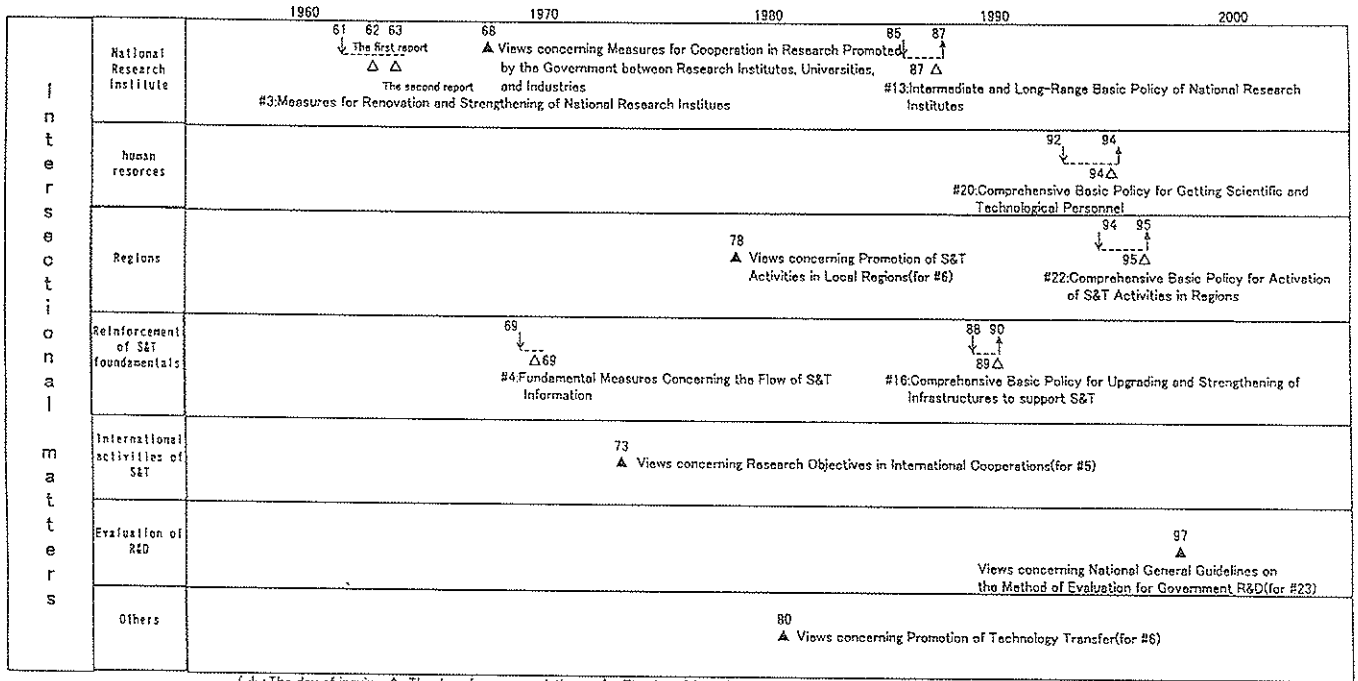


Table 6. *Changes in the Framework of Japanese Science and Technology Policy.*

1956	Science & Technology Agency (STA) was established to promote Japan's S&T (except for research in universities).
1959	Council for Science & Technology (CST) was established to comprehensively formulate and coordinate overall science & technology policy for the Prime Minister.
1960	*CST produced the Recommendation for Inquiry No. 1 "Comprehensive and Fundamental Measures for the Development of Science and Technology for the next 10 years".*
1971	*CST produced the Recommendation for Inquiry No. 5 "The Fundamentals of Comprehensive S&T policy for the 1970s".*
1977	*CST produced the Recommendation for Inquiry No. 6 "The Foundation of Japan's Over All Science and Technology Policy Based on Long-Term Prospects".*
1981	"Special Coordination Fund for Promoting Science and Technology" was established to facilitate the comprehensive promotion of S&T.
1982	'Provisional Research Council for Administration' recommended strengthening the coordination function in S&T administration.
1983	'Committee on Policy Matters' was established within CST.
1984	*CST provided the Recommendation for Inquiry No. 11 "Comprehensive Fundamental Policy for Promotion of Science and Technology according to the Changing Situations from the Long Term Pomt of View.*
1985	'Provisional Council for Promotion of Administrative Reform' recommended the establishment of a General Guideline for Science & Technology Policy and the reform of some bureaux in STA.
1986	*"General Guideline for Science & Technology Policy" based on the recommendation of CST was approved at the Cabinet Meeting.*
	Three bureaux in STA were reformed.
	The Law for Facilitating Government Research Exchange was enacted.
1988	"National Institute of S&T Policy (NISTEP)" was founded as a core institute to conduct theoretical and survey-based research on S&T Policy.
1992	*CST produced the Recommendation for Inquiry No. 18, "Comprehensive and Basic Science and Technology Policy toward the New Century".*
	"General Guideline for Science & Technology Policy" based on the above recommendation was approved at the Cabinet Meeting.
1995	"Science & Technology Basic Law" was enacted.
1996	*"Science & Technology Basic Plan" was approved at the Cabinet Meeting after consultation with CST.*
1997	*The National Guideline on the Method of Evaluation for Government R&D drafted by CST, was approved by Prime Minister.*
1998	The Administrative Reform Basic Law was enacted.

Events related to the recommendation of CST are in Italics.

tions on comprehensive policy, the CST has made more specific policy recommendations in response to, and in anticipation of, the needs of society as a whole. All recommendations made by the CST are listed in table 6.

Thus, the CST has played a pivotal role in the evolution of S&T policy since its foundation four decades ago. The development of Japanese S&T policy is summarized in table 4. Although the framework for S&T policies has greatly changed, the administrative structure for the formulation and the implementation of S&T policy has remained largely untouched. The legal status and functions of the CST and related ministries and agencies have not greatly changed until today. At present, however, the concrete role and organization of the new General Council for Science and Technology (GCST) and its secretariat are being discussed in connection with various S&T related organizations whose reforms are expected to start in the year 2001.

Conclusion

The change in the framework for S&T policy in Japan is inexorable and intrinsically intertwined with the change in the entire Japanese socio-economic system within a dynamic global environment. The direction of change is one in which various functions involved in the formulation and the execution of S&T policy are becoming more distinctive and transparent and require a more contract-oriented society. A concrete plan for administrative reform has been elaborated in the field of science and technology as well as in other fields of governmental services.

Mere reform without a clearly thought out and acceptable design for implementing the new model for S&T policy could make the situation worse instead of building upon the positive aspects of the existing system. Good design of reform processes is only possible if the characteristics of the major actors involved and their interactions are properly understood. The recent changes described at the outset of this paper have initiated a movement toward a radical reform in the framework for S&T policy. However, the actual adoption of new practices and new modes of thinking in the formulation and the implementation of S&T policy are rare and still tentative.

Under these circumstances, great efforts are needed to change the Japanese S&T system so that it can become effective and attractive in a highly competitive world. The profound change Japan is facing could turn out to be a chance for a radically new system for research, development and innovation. This new system is characterized by flexibility, competition, and a more open reseach environment. The emphasis is on creativity and responsiveness to social and economic needs, as well as responsiveness to

opportunities arising from scientific discovery itself. The emergence of a knowledge-based society in the twenty-first century will surely challenge Japan to base its source of vitality on science and technology in the broadest possible perspective.

REFERENCES

Fujigaki, Y., and Nagata, A. (1998): *Concept Evolution, Science and Public Policy.*

Government of Japan (1996): *Science and Technology Basic Plan* (Tokyo, Science and Technology Agency).

Kawasaki, M. (1994): 'Historical Development of Japanese Science and Technology Policy in Conjunction with socio-economic Policy', in G. Aichholzer and G. Schieustock (eds.), *Technology Policy: Towards or Integration of Social and Ecological Concerns.*

Mori, W., and Ochiai, T. (1997): 'Science and Technology in Japan', in Golden, W.T. and Ratchford, J.T. (eds.), *Technology in Society*, Vol. 19, No. 3/4.

National Institute of Science and Technology Policy (1997): *Science and Technology Indicators: 1997* (Tokyo, National Institute of Science and Technology Policy, Science and Technology Agency).

Sato, Y. (1998): 'The Changing Framework for the S&T Policy in Japan', Presentation to NISTEP 10th Anniversary Conference, Tokyo, October 8-9, 1998.

Science and Technology Agency, *White Paper on Science and Technology*, annual editions.

Finito di stampare nel mese di Luglio 2000
dalla Tipografia della Pace
Via degli Acquasparta, 25 – 00186 Roma