PONTIFICIA
ACADEMIA
SCIENTIARVM

Acta
19

Predictability in Science: Accuracy and Limitations

Edited by

Werner Arber
Nicola Cabibbo
Marcelo Sánchez Sorondo

The Proceedings
of the Plenary Session
3-6 November 2006

VATICAN CITY
2008

PREDICTABILITY IN SCIENCE: ACCURACY AND LIMITATIONS

Address:

The Pontifical Academy of Sciences
Casina Pio IV, 00120 Vatican City
Tel: +39 0669883195 / Fax: +39 0669885218
Email: academy.sciences@acdscience.va

PONTIFICIAE ACADEMIAE SCIENTIARVM ACTA

19

The Proceedings
of the Plenary Session on

PREDICTABILITY IN SCIENCE: ACCURACY AND LIMITATIONS

3-6 November 2006

Edited by
Werner Arber
Nicola Cabibbo
Marcelo Sánchez Sorondo

EX AEDIBVS ACADEMICIS IN CIVITATE VATICANA

MMVIII

The opinions expressed with absolute freedom during the presentation of the papers of this meeting, although published by the Academy, represent only the points of view of the participants and not those of the Academy.

ISBN 978-88-7761-094-2

PONTIFICIA ACADEMIA SCIENTIARVM
VATICAN CITY

His Holiness Pope Benedict XVI

The Participants of the Plenary Session of 3-6 November 2006

Papal Audience of 5 November 2006

The Participants of the Plenary Session of 3-6 November 2006

CONTENTS

SCIENTIFIC PAPERS

Session I – Chaos and Predictions in Physics and Astronomy

Session II – Geosciences and Environmental Evolution

Session III – Predictions in the Life Sciences

Session IV – Philosophical and Societal Aspects

Session V – Research Procedures: Theories and Their Verification, Serendipity

Session VI – Public Perception and Policy in the Context of Uncertainty

INTRODUCTION

In its last two business meetings the PAS Council discussed several alternative topics that had been proposed for the next Plenary Session of the Pontifical Academy of Sciences in November 2006. On the basis of these deliberations the Council has chosen the theme of 'Predictability in Science'. This theme is situated at the interphase between fundamental science and its practical applications to the benefit of human beings. We will welcome contributions on scientific predictions of impending dangers, such as earthquakes, on the outlook for climate change, on the analysis of nuclear and other technologies, on the role of prediction in the medical sciences, and on many other scientific predictions and modelling approaches that frequently also have their applications with impact on cultural and socio-political developments.

According to Webster's dictionary the word *prediction* has two meanings: (1) a predicting or being predicted, and (2) a prophecy. Obviously, only the first meaning applies to our proposed theme. Therefore the word science in the title is important. It is not our role to debate on prophecies. Predictions on a scientific basis can be more or less accurate and have in most cases their intrinsic limitations. We therefore consider expressing this in a subheading. The proposed theme would then read as follows: 'Predictability in Science: Accuracy and Limitations of Predictions based on Scientific Knowledge'.

We expect that all scientific disciplines can contribute with selected examples to a wide debate on scientific predictions and their relevance to society. Thereby, the awareness of natural limitations that are inherent to many predictions plays an important role and it can ensure the trust in science in interactions between science and the civil society. The distinction between certainties and uncertainties has been made for a long time by many scientists presenting scientific knowledge, theories and models on natural developments. A candid discussion on this theme by the Pontifical Academy of Sciences can represent a clarifying view for both the scientific community and the general public.

Werner Arber & Nicola Cabibbo

BACKGROUND NOTE

The idea behind this meeting is that a great deal is done within science that only scientists would deal with; but there are, increasingly, many other aspects of science which have public implications that figure extensively in the audiovisual and print media. With regard to the latter, a public understanding of accuracy and uncertainty in the predictions, arising from scientific knowledge, becomes important.

Most of these aspects of general public interest relate to complex phenomena; and some relate to ethical issues that could have political, social and economic ramifications. For example, issues relating to weather prediction, climate change, prediction of earthquakes, the possibility of major natural disasters such as an asteroid hitting the earth, pandemics from the SARS virus, bird flu and the like that could cross over from animals to people, all figure in societal discussions these days. In most of these cases, the issue is that whilst there is a very good scientific basis at the initial stages, as one proceeds down the line, to predictions of relevance to society, many aspects such as complexity, non-equilibrium phenomena, chaos and the like come in, introducing uncertainties in the predictions.

Thus, one has clear-cut observations on the carbon dioxide concentrations in the atmosphere and how they have increased with time. One knows that carbon dioxide is a greenhouse gas; and there are also other greenhouse gases. The increase in their concentrations will result in a greenhouse effect, which will cause a rise in global temperature. Sources, pathways, sinks and budgets at each stage will define what ultimately happens.

These will have implications on sea levels, change in climate patterns, extreme events in precipitation, availability of water etc. At each stage there is a greater degree of uncertainty, and an increasingly poorer predictability and lack of consensus. Society gets confused and begins to doubt whether scientists know what they are talking about e.g. the fun many have at the expense of meteorologists. When one gets to the human dimensions of global change, behavioural change etc come into the picture, e.g involving economics, psychology and social behaviour, and human dislike for measures that demand lifestyle changes. This is the type of problem dealt with by the Inter-Governmental Panel on Climate Change.

The above somewhat lengthy real life example was only to illustrate the nature of the problem because the public generally assumes that science can be certain and give accurate answers.

There is then the possibility of climate engineering to reduce greenhouse effects. But this could raise ethical issues related to tampering with a natural system and predictability of the hazards involved in this.

There is the issue relating to genetically modified organisms and their large-scale use in society. There are many who are opposed to this on the grounds that one would not know what might ultimately happen; and more particularly in relation to the environment.

There is the broad area of environment and ecology where society seldom looks at the price being paid for certain pathways of development. For example, what is the price to be attached to the ecological services provided by water?

Today, there are very significant advances in modelling capabilities and one needs to evaluate the accuracy and limitations in prediction based on these techniques. While it would be possible with increasing knowledge and capabilities to have predictions in many areas, these would never be 100% correct, and one would have to live with uncertainty. In fact, it is this uncertainty that makes further development of science exciting, because there is so much more to know and to understand.

However, there is also the question of decisions that have to be taken by governments and society at any point in time for which advice from the scientific community is called for. This would bring out the importance of the precautionary principle, to avoid getting into a situation that might lead to catastrophic events.

It is felt that the intrinsic issue of scientific uncertainty, particularly in complex, non-equilibrium systems, and limits of predictability need to be discussed from the viewpoint of various angles. It is felt that apart from natural scientists it would be important to have some distinguished thinkers who deal with dimensions that human society is normally concerned with e.g. economic, social and behavioural aspects also participate in this plenary session. This is because many of these areas are characterised by non-equilibrium complex situations and are also increasingly using the techniques developed in the pure sciences for their analysis.

M. Govind Kumar Menon

PROGRAMME

FRIDAY, 3 NOVEMBER 2006

9:00 Welcome
Prof. Nicola Cabibbo, President of the Pontifical Academy of Sciences

9:05 The Subject of the Meeting
Prof. Werner Arber, Coordinator of the Meeting and PAS Academician

9:15 Commemoration of Prof. Richard Southwood by Prof. Raymond Hide

9:30 Self-Presentation of Prof. Theodor W. Hänsch, Prof. Ingo Potrykus

10:15 Coffee Break

SESSION I
Chaos and Predictions in Physics and Astronomy
Chairperson: Prof. Werner Arber

10:45 Prof. Rudolf Muradian
Predictions in Astrophysics and Cosmology
Discussion

11:30 Prof. Antonino Zichichi
Complexity and Predictions at the Fundamental Level of Scientific Knowledge
Discussion

12:30 Lunch at the Casina Pio IV

SESSION II
Geosciences and Environmental Evolution
Chairperson: Prof. Nicole M. Le Douarin

15:00 Prof. Vladimir Keilis-Borok
Predictability of Complex Systems, with Special Reference to Geological Disasters
Discussion

16:00 Prof. Veerabhadran Ramanathan
Global Warming Science: Predictions, Surprises and Insurmountable Uncertainties
Discussion

17:00 Coffee Break

17:30 Prof. Mario J. Molina
Predictability of Science and Climate Change
Discussion

18:30 Prof. Paul J. Crutzen
An Example of Geo-Engineering: Cooling Down Earth's Climate by Sulfur Emissions in the Stratosphere
Discussion

19:30 Dinner at the Casina Pio IV

Saturday, 4 November 2006

Session III
Predictions in the Life Sciences
Chairperson: Prof. William D. Phillips

9:00 Prof. Rafael Vicuña
Attempts to Predict a Minimal Genome
Discussion

10:00 Prof. Umberto Veronesi
The New Possibilities of Prediction and Prevention of Cancer
Discussion

11:00 Coffee Break

11:30 Prof. Werner Arber
Stochastic Genetic Variations and their Role in Biological Evolution
Discussion

12:30 Lunch at the Casina Pio IV

Session IV
Philosophical and Societal Aspects
Chairperson: Prof. M. Govind Kumar Menon

14:00 Prof. Jean-Michel Maldamé
Epistemological Study of the Vocabulary of Prediction in Science and in Theology
Discussion

15:00 Prof. Michael Heller
Predictability, Measurement and Cosmic Time
Discussion

16:00 Coffee Break

16:30 Prof. Jürgen Mittelstrass
Epistemological Remarks on the Concept of Predictability
Discussion

17:30 Prof. Antonio Battro
Predictability: Prophecy, Prognosis and Prediction. A Study in Neuroeducation
Discussion

18:30 Dinner at the Casina Pio IV

SUNDAY, 5 NOVEMBER 2006

Visit to the Museo Nazionale Romano, Palazzo Massimo, Rome
Presentation of the Pius XI Medal to Prof. Ashoke Sen
Lunch at the Casina Pio IV

MONDAY, 6 NOVEMBER 2006

SESSION V
Research Procedures: Theories and their Verification, Serendipity
Chairperson: Prof. Paul J. Crutzen

9:00 Prof. William D. Phillips
When Results are Better than Predicted: A Case Study
Discussion

10:00 Prof. Michael Sela
On Unpredictability in Research Projects
Discussion

11:00 Fr. Prof. Stanley L. Jaki
Science as Prediction and the Unpredictability of Science
Discussion

11:30 Audience with the Holy Father Pope Benedict XVI

13:30 Lunch at the Casina Pio IV

Session VI
Public Perception and Policy in the Context of Uncertainty
Chairperson: Prof. Nicola Cabibbo

15:00 Prof. M. Govind Kumar Menon
A Short Background Note
Discussion

16:00 General Discussion

16:45 Coffee Break

17:30 Closed Session for Academicians

18:30 Dinner at the Casina Pio IV

LIST OF PARTICIPANTS

Prof. Nicola CABIBBO, President
H.E. Msgr. Prof. Marcelo Sánchez Sorondo, Chancellor
The Pontifical Academy of Sciences
Casina Pio IV
00120 Vatican City

Academicians

Prof. Werner Arber
University of Basel
Department of Microbiology
Biozentrum
Klingelbergstrasse 70
CH-4056 Basel (Switzerland)

Prof. Antonio M. Battro
Battro & Denham
Billinghurst 2574 Piso 1 A
C1425DTZ Buenos Aires (Argentina)

Prof. Enrico Berti
Università degli Studi di Padova
Dipartimento di Filosofia
Piazza Capitaniato, 3
I-35139 Padova (Italy)

Prof. Bernardo M. Colombo
Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Via Battisti, 241
I-35121 Padova (Italy)

H.Em. George Card. Cottier
Santa Marta
00120 Vatican City

Prof. Paul J. Crutzen
Max-Planck-Institute For Chemistry
Department of Atmospheric Chemistry
P.O. Box 3060
D-55020 Mainz (Federal Republic of Germany)

Prof. Albert Eschenmoser
ETH Hönggerberg HCI H309
Laboratorium für Organische Chemie
Wolfgang-Pauli-Strasse 10
CH-8093 Zürich (Switzerland)

Prof. Theodor Hänsch
Max-Planck-Institut Für Quantenoptik
Hans-Kopfermann-Strasse, 1
D-85748 Garching (Federal Republic of Germany)

Rev. Prof. Michael Heller
ul. Powstancow Warszawy, 13/94
PL-33-110 Tarnow (Poland)

Rev. Prof. Stanley L. Jaki
P.O. Box 167
Princeton, N.J. 08542 (U.S.A.)

Prof. Vladimir I. Keilis-Borok
University Of California, Los Angeles
Institute of Geophysics and
Planetary Physics
3845 Slichter Hall, Box 951567 - Of. 1813a Geology Building
Los Angeles, CA 90095-1567 (U.S.A.)

Prof. Nicole Le Douarin
Académie des sciences
23, quai de Conti
F-75006 Paris (France)

Prof. Rita Levi-Montalcini
European Brain Research Institute (EBRI)
Via del Fosso di Fiorano, 64/65
I-00143 Roma (Italy)

Rev. Prof. Jean-Michel Maldamé
Couvent Saint-Thomas d'Aquin
Impasse Lacordaire
F-31078 Toulouse Cedex 4 (France)

Prof. M. Govind Kumar Menon
C-178 (FF), Sarvodaya Enclave
New Delhi 110017 (India)

Prof. Jürgen Mittelstrass
Konstanz Universität
Philosophische Fakultät
Fachgruppe Philosophie
Postfach 5560 D 15
D-78434 Konstanz (Federal Republic of Germany)

Prof. Mario J. Molina
University of California, San Diego
Department of Chemistry and Biochemistry
2040 Urey Hall Addition
9500 Gilman Drive, MC 0356
La Jolla, CA 92093 (U.S.A.)

Prof. Rudolf Muradian
Rua Ezequiel Ponde, 106, apt. 1002 - Jardim Apipema
40155-050 Salvador BA (Brazil)

Prof. Crodowaldo Pavan
Rua Alvares Florence, 298 (Butantan)
05502-060 Sao Paulo, S.P. (Brazil)

Prof. William D. Phillips
National Institute of Standards and Technology
Building 216, Room B133
Gaithersburg, MD 20899 (U.S.A.)

Prof. Ingo Potrykus
Im Stigler 54
CH-4312 Magden (Switzerland)

Prof. Veerabhadran Ramanathan
University of California, San Diego
Scripps Institution of Oceanography
Center for Atmospheric Sciences
9500 Gilman Drive, MC 0221
La Jolla, CA 92093-0221 (U.S.A.)

Prof. Chintamani N.R. Rao
Jawaharlal Nehru Centre for Advanced Scientific Research
Jakkur Post
Bangalore 560064 (India)

Prof. Michael Sela
The Weizmann Institute Of Science
Department of Immunology
P.O. Box 26
Rehovot 76100 (Israel)

Prof. Wolf J. Singer
Max-Planck-Institute for Brain Research
Department of Neurophysiology
Deutschordenstrasse 46
D-60528 Frankfurt Am Main (Federal Republic of Germany)

Prof. Walter E. Thirring
Universität Wien
Institut für Theoretische Physik
Boltzmanngasse 5
A-1090 Vienna (Austria)

Prof. Charles H. Townes
University of California, Berkeley
Department of Physics
366 LeConte Hall, # 7300
Berkeley, CA 94720-7300 (U.S.A.)

Prof. Hans Tuppy
University of Vienna
Institute of Biochemistry
Dr. Bohr-Gasse 9,3.Stock
A-1030 Vienna (Austria)

Prof. Rafael Vicuña
Pontificia Universidad Católica de Chile
Departamento de Genetica Molecolar y Microbiología
Casilla 114-D
Santiago (Chile)

Prof. Antonino Zichichi
European Organization for Particle Physics (Cern)
CH-1211 Geneva 23 (Switzerland)

Pius XI Medal

Prof. Ashoke Sen
Harish-Chandra Research Institute
Chhatnag Riad, Jhusi
Allahabad 211019 (India)

Expert

Prof. Umberto Veronesi
Istituto Europeo di Oncologia
Via Ripamonti, 435
20141 Milano (Italy)

ADDRESS OF THE PRESIDENT TO THE HOLY FATHER

Holy Father,

Meeting you once more fills us with joy. We have followed your recent Magisterium where you often quoted with admiration the first leader of our Academy, Galileo Galilei, and said, 'The positive aspects of modernity are to be acknowledged: we are all grateful for the marvellous possibilities that it has opened up for mankind'.[1] It is these positive aspects of modernity, which science so generously offers us, that are the focus of our current meeting on 'Predictability in Science. Accuracy and Limitations'. We come from different regions of the world: many of us Christians, others Jewish and from other religions. However, we all agree that, now more than ever, we need to keep in mind your statement that 'The scientific ethos is the will to be obedient to the truth, and, as such, it embodies an attitude which belongs to the essence of the Christian spirit'.[2] Indeed, science was born to search for the truth of the 'natural world' and, in doing so, frees itself from the conditionings of power and manipulation.

This year the Academy is welcoming four new members: Prof. Ingo Potrykus, Plant Biologist, appointed on 10 March 2005; Prof. Theodor Hänsch, Physicist, Nobel Prize in Physics in 2005, appointed on 15 May 2006; Prof. Edward Witten, Physicist and Mathematician, Fields Medal in 1990, appointed on 15 May 2006; Prof. José Funes, Astronomer, Director of the Vatican Specola, appointed on 5 August 2006.

The Academy would like to take this opportunity to thank you for the appointment of these new members and for your constant solicitude in its regard, and wants to renew its special relationship with Pope Benedict XVI, whom we consider one of our most distinguished members, not only, of course, as Pope, but also for your intellectual achievements.

[1] *Lecture at the Aula Magna of the University of Regensburg*, 12 September 2006.

[2] *Ibid.*

We believe you will be pleased that, for our next Plenary Session, the Council has decided to study a subject that we know is of your special interest: 'Scientific insights into the evolution of the Universe and of Life'.

The Academy agrees that to succeed in overcoming the dangers arising from the new possibilities that science offers to humanity, faith and reason need to come together in a new way. This encounter must overcome the self-imposed limitations of scientific reason to the empirically falsifiable, while philosophical and theological reason must listen more attentively to scientific reason. We are grateful for the attention you have devoted to this encounter of the religious and philosophical world with our scientific world and we are striving to meet your high expectations.

Thank you.

Nicola Cabibbo

ADDRESS OF HIS HOLINESS BENEDICT XVI TO THE MEMBERS OF THE PONTIFICAL ACADEMY OF SCIENCES

Monday, 6 November 2006

Your Excellencies,
Distinguished Ladies and Gentlemen,

I am pleased to greet the members of Pontifical Academy of Sciences on the occasion of this Plenary Assembly, and I thank Professor Nicola Cabibbo for his kind words of greeting in your name. The theme of your meeting – 'Predictability in Science: Accuracy and Limitations' – deals with a distinctive attribute of modern science. Predictability, in fact, is one of the chief reasons for science's prestige in contemporary society. The establishment of the scientific method has given the sciences the ability to predict phenomena, to study their development, and thus to control the environment in which man lives.

This increasing 'advance' of science, and especially its capacity to master nature through technology, has at times been linked to a corresponding 'retreat' of philosophy, of religion, and even of the Christian faith. Indeed, some have seen in the progress of modern science and technology one of the main causes of secularization and materialism: why invoke God's control over these phenomena when science has shown itself capable of doing the same thing? Certainly the Church acknowledges that 'with the help of science and technology…, man has extended his mastery over almost the whole of nature', and thus 'he now produces by his own enterprise benefits once looked for from heavenly powers' (*Gaudium et Spes*, 33). At the same time, Christianity does not posit an inevitable conflict between supernatural faith and scientific progress. The very starting-point of Biblical revelation is the affirmation that God created human beings, endowed them with reason, and set them over all the creatures of the earth. In this way, man has become the steward of creation and God's 'helper'. If we think, for example, of how modern science, by predicting natural phenomena, has contributed to the protection of the environment, the progress of developing nations, the fight against epidemics, and an increase in life expectancy, it becomes clear that there is no conflict between God's providence and human enterprise. Indeed, we could say that the work of predicting, controlling and governing nature, which science today renders more practicable than in the past, is itself a part of the Creator's plan.

Science, however, while giving generously, gives only what it is meant to give. Man cannot place in science and technology so radical and unconditional a trust as to believe that scientific and technological progress can explain everything and completely fulfil all his existential and spiritual needs. Science cannot replace philosophy and revelation by giving an exhaustive answer to man's most radical questions: questions about the meaning of living and dying, about ultimate values, and about the nature of progress itself. For this reason, the Second Vatican Council, after acknowledging the benefits gained by scientific advances, pointed out that the 'scientific methods of investigation can be unjustifiably taken as the supreme norm for arriving at truth', and added that 'there is a danger that man, trusting too much in the discoveries of today, may think that he is sufficient unto himself and no longer seek the higher values' (*ibid.*, 57).

Scientific predictability also raises the question of the scientist's ethical responsibilities. His conclusions must be guided by respect for truth and an honest acknowledgment of both the accuracy and the inevitable limitations of the scientific method. Certainly this means avoiding needlessly alarming predictions when these are not supported by sufficient data or exceed science's actual ability to predict. But it also means avoiding the opposite, namely a silence, born of fear, in the face of genuine problems. The influence of scientists in shaping public opinion on the basis of their knowledge is too important to be undermined by undue haste or the pursuit of superficial publicity. As my predecessor, Pope John Paul II, once observed: 'Scientists, precisely because they "know more", are called to "serve more". Since the freedom they enjoy in research gives them access to specialized knowledge, they have the responsibility of using that knowledge wisely for the benefit of the entire human family' (*Address to the Pontifical Academy of Sciences,* 11 November 2002).

Dear Academicians, our world continues to look to you and your colleagues for a clear understanding of the possible consequences of many important natural phenomena. I think, for example, of the continuing threats to the environment which are affecting whole peoples, and the urgent need to discover safe, alternative energy sources available to all. Scientists will find support from the Church in their efforts to confront these issues, since the Church has received from her divine founder the task of guiding people's consciences towards goodness, solidarity and peace. Precisely for this reason she feels in duty bound to insist that science's ability to predict and control must never be employed against human life and its dignity, but always placed at its service, at the service of this and future generations.

There is one final reflection that the subject of your Assembly can suggest to us today. As some of the papers presented in the last few days have emphasized, the scientific method itself, in its gathering of data and in the processing and use of those data in projections, has inherent limitations that necessarily restrict scientific predictability to specific contexts and approaches. Science cannot, therefore, presume to provide a complete, deterministic representation of our future and of the development of every phenomenon that it studies. Philosophy and theology might make an important contribution to this fundamentally epistemological question by, for example, helping the empirical sciences to recognize a difference between the mathematical inability to predict certain events and the validity of the principle of causality, or between scientific indeterminism or contingency (randomness) and causality on the philosophical level, or, more radically, between evolution as the origin of a succession in space and time, and creation as the ultimate origin of participated being in essential Being.

At the same time, there is a higher level that necessarily transcends all scientific predictions, namely, the human world of freedom and history. Whereas the physical cosmos can have its own spatial-temporal development, only humanity, strictly speaking, has a history, the history of its freedom. Freedom, like reason, is a precious part of God's image within us, and it can never be reduced to a deterministic analysis. Its transcendence vis-à-vis the material world must be acknowledged and respected, since it is a sign of our human dignity. Denying that transcendence in the name of a supposed absolute ability of the scientific method to predict and condition the human world would involve the loss of what is human in man, and, by failing to recognize his uniqueness and transcendence, could dangerously open the door to his exploitation.

Dear friends, as I conclude these reflections, I once more assure you of my close interest in the activities of this Pontifical Academy and of my prayers for you and your families. Upon all of you I invoke Almighty God's blessings of wisdom, joy and peace.

COMMEMORATION OF DECEASED ACADEMICIAN

T. Richard Edmund Southwood († 26.X.05)

Emeritus Linacre Professor of Zoology and former Vice-Chancellor of the University of Oxford, Sir Richard Southwood (hereafter RS) died in Oxford on 26 October 2005. By his own account (see his entry in the 2004 Yearbook of the Pontifical Academy of Sciences), his main scientific contributions were in ecology with a strong bias to entomology, and environmental sciences and policy. A Fellow of the Royal Society, during his distinguished career he received many other honours in recognition not only of his influential research and teaching in zoology but also of his activities as an outstanding university administrator and valued government advisor. He became a member of the Pontifical Academy of Sciences in 1992.

RS was born on 20 June 1931 in the town of Northfleet in the County of Kent in south-east England, where his father owned a dairy farm, and he received his secondary education at the nearby Gravesend Grammar School, which he attended from 1942-49. He then went on to Imperial College London where in 1955 he obtained a PhD degree on the basis of research on time trends and patterns of species diversity, making use of the long-term data sets on insects held at the Rothamsted Experimental Station near Harpenden, where he met and married his future wife Alison Langley.

During his subsequent career, he served first on the staff of Imperial College, where he became Professor of Zoology and Applied Entomology and Chairman of the Division of Life Sciences, and then, from 1979-93, as Linacre Professor of Zoology at the University of Oxford. His outstanding skills as an administrator from which Imperial College London and the Department of Zoology at Oxford University had already greatly benefited led to his appointment as Vice-Chancellor of the University (from 1989-93). A Vice-Chancellor's many duties include fund-raising on behalf of the University, another activity in which RS enjoyed notable success.

Outside the University RS served with distinction as chairman of an international conference held in 1986 on Biological Effects of Low-Level

Radiation and also of several U.K. government bodies also dealing with important and politically sensitive issues, including the Royal Commission on Environmental Pollution (1981-86), National Radiological Protection Board (1985-94), Working Party on Bovine Spongiform Encephalopathy (1988-89), Round Table on Sustainable Development (1995-99) and Interagency Committee on Global Environmental Change (1997-2000).

According to one close colleague, RS was 'one of the most notable ecologists and zoologists of his generation (but) he leaves an even more lasting legacy through his superb skills as a mentor and builder of academic departments – first at Imperial College London, and later at Oxford University – whose distinguished individuals added up to more than the sum of their parts. A disproportionate number of the world's top ecological researchers today are British, and almost all of them were directly influenced by him'.

Notwithstanding his many other duties at Oxford, for eighteen years he found the time to give stimulating undergraduate lectures there. These provided the basis of his last book *The Story of Life,* published by the Oxford University Press in 2002 (paperback 2003). The excerpts from this 'masterly overview impressive in depth, breadth and clarity of the origin and evolution of life' that were selected for presentation at a moving memorial service held last February in the Chapel of Merton College Oxford to celebrate the life of this remarkable man were the opening and closing paragraphs of the book. For their eloquence and the challenges they present to all of us they are worth repeating here.

'Consider the amazing variety of life today: the great herds of animals that roam the African plains, the shoals of fish that teem in coral reefs or the flocks of penguins that huddle on the Antarctic ice. Yet what we see around us is but one still from the film ("movie") of life, a glimpse that we can only understand if we know what came before. This is the book of the film of all life'.

'Will humans having made so much progress by increasing the carrying capacity of their habitat finally end by overexploiting the world and giving the kaleidoscope another shake? But life is flexible, and we can be sure that the frame of the kaleidoscope will be filled with a new pattern of colours. In contrast, we, in our prodigious numbers, are locked by our agricultural and commercial activities into the current climatic regime. Can political stability survive the stresses that will arise when this changes or will we doom ourselves? We carry a burden of responsibility to learn from our knowledge of the world and its past. Time is short, but we do have the ability to change'.

Raymond Hide

SELF-PRESENTATIONS OF THE NEW MEMBERS

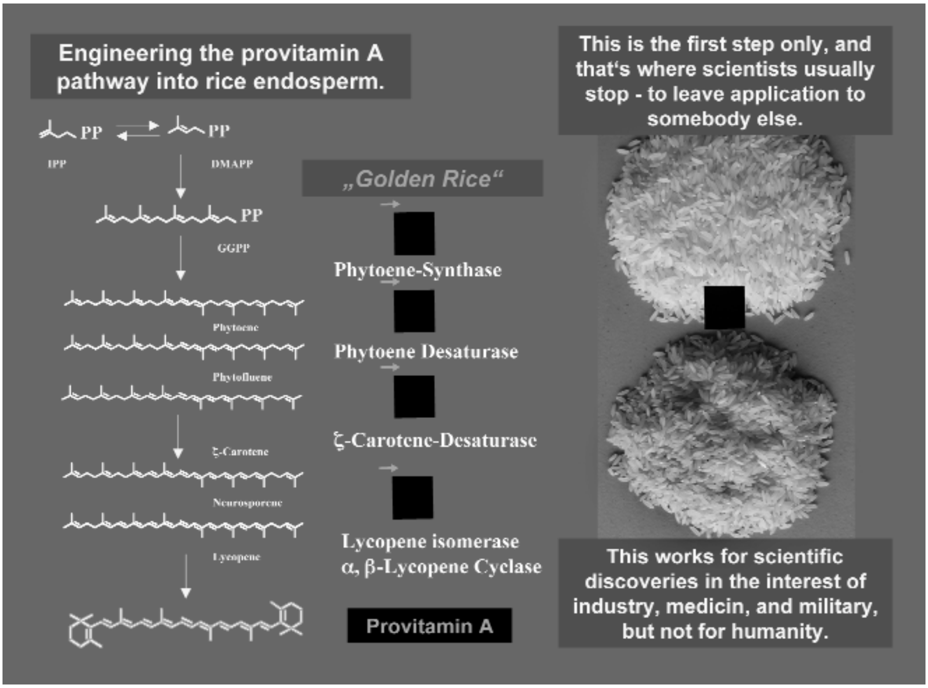
Ingo Potrykus

Born into the family of Dr. med. Waldemar and Hildegard Potrykus on 5 December 1933 in Hirschberg, Silesia, Germany, I had a peaceful youth with my two brothers until we escaped from the Russian army in February 1945. Our life changed dramatically when we had to survive as refugees in Northern Bavaria, we lost our father a few days before the war was over and one day after our sister was born. We had lost everything and it was only thanks to the admirable persistence of our mother that we could all complete higher education. I met my wife early in our life. She was 16 and I was 19, and since then we have been together, having now a family of three children and eight grandchildren. Already then my life was deeply influenced by my attraction to traditional biology and I finished my studies as college teacher in biology.

My conversion to real science was not planned and was initiated by Prof. Josef Straub, director of the Max-Planck-Institute for Plant Breeding Research Cologne, who offered me the chance to work for a PhD in plant genetics. The work for this thesis got me fascinated in the biological phenomenon of 'totipotency' of somatic plant cells (somatic plant cells often have the potential to grow to a complete plant, when isolated and provided with the appropriate nutritional environment). This was towards the end of the 60s and at the peak of the 'Green Revolution'. Probably because of my experience as a refugee, I was already concerned about food security of the poor in developing countries, and I had the impression that this required more support. Being more of an 'engineer' than a 'scientist', I was intrigued to challenge the potential offered by the phenomenon of totipotency for an improved food security.

The director of the Institute of Plant Physiology, University of Hohenheim, Professor Dieter Hess gave me the opportunity to begin a university career in 1970 by working on plant tissue culture and genetic engineering. This opportunity was amplified when the chairman of the Max-Planck-Institute of Plant Genetics Heidelberg, Professor Georg Melchers offered me the position of a research group leader in 1974. Conditions to work towards technology development for genetic engineering of crop plants further increased with my appointment to the Friedrich Miescher Institute in Basel 1976, and they reached the optimum with my appointment as full professor in plant sciences at the Institute of Plant Sciences of the Swiss Federal Institute of Technology (ETH) Zürich. This new institute, which I established together with my colleague Professor Josef Nösberger, was the ideal environment for my intentions: Three professors in basics plant biology, three in agronomy, and three in plant protection joined in one institute – a concept developed under the leadership of the Academy's long-standing member Professor Werner Arber.

My interest in science is that of a 'tool' to help solve humanitarian problems: e.g. plant molecular biology and cell biology are for me the basis of plant biotechnology, and plant biotechnology is for me a tool for improved food security of the poor in developing countries. The case of 'Golden Rice' exemplifies this philosophy: Vitamin A-malnutrition takes a daily toll of 6,000 lives. I assumed that rice engineered to contain provitamin A would be a cost-effective and sustainable intervention to reduce vitamin A-malnutrition. This



Engineering the provitamin A pathway into rice endosperm.

This is the first step only, and that's where scientists usually stop - to leave application to somebody else.

„Golden Rice"

Phytoene-Synthase

Phytoene Desaturase

ζ-Carotene-Desaturase

Lycopene isomerase
α, β-Lycopene Cyclase

Provitamin A

This works for scientific discoveries in the interest of industry, medicin, and military, but not for humanity.

goal required the development of state-of-the-art science and technology for the crop plant rice, not just for a model plant as most of my colleagues were focusing on. And it required above all something scientists normally try to avoid – reaching out far beyond the 'ivory tower of science'.

When I proposed the project to engineer the biochemical pathway for the synthesis of provitamin A (plants do not produce vitamin A, but provitamin A and the human body converts it to vitamin A in a carefully controlled reaction), the scientific community considered this project (rightly) totally unfeasible for biological and technical reasons. It was, therefore, difficult to attract the necessary funding. Thanks to my privileged position as professor at the ETH Zürich, I could use my own funds to start a PhD thesis on it. I approached The Rockefeller Foundation for complementing funding which, following a brainstorming in New York in 1992, confirming the extremely low chances for success, decided to do so, because in case of success, the outcome would have such a high potential for reducing vitamin A-malnutrition. We were fortunate and after nearly ten years of experimentation and a final co-transformation experiment with five genes, we could present a yellow rice synthesizing and accumulating provitamin A to the public. This was 31 March 1999 at my Farewell Symposium from the ETH, which I had to leave because I had passed the age of 65. Golden Rice was the result of a perfect collaboration of my team with the team of Dr. Peter Beyer from the university of Freiburg, Germany. The scientific community, the media, and the public were excited and there was lot of recognition.

I was determined that this was not the end of our project. However, we soon realized that the public sector was not prepared for any continuation beyond basic science.

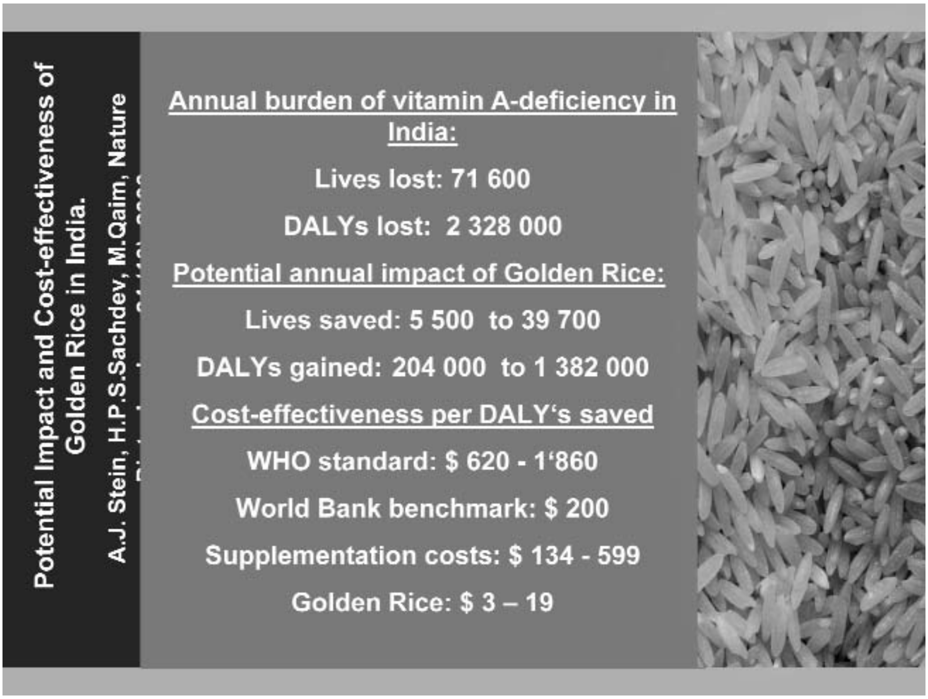
Humanitarian application is not on the agenda of the public domain nor on that of public funding systems. And solutions for humanitarian problems are, of course, not under the responsibility of the private sector. 'Golden Rice' would have remained just a scientific curiosity and would not have saved a single child, if we had stayed within the 'ivory tower'. To act responsibly we had to move into many new areas and had – and have – to fight many unpleasant battles. We had to acquire free licenses for humanitarian use for all intellectual property rights involved. We had to get access to know-how and financial support for product development and deregulation. We had to adjust our GMO-events to regulatory requirements. We had to find competent partner institutions in developing countries. We had (and still have) to develop agronomical competitive national varieties. We had to defend the project against aggressive anti-GMO lobbying and we had to try to provide the media and the public with correct information, to respond to the growing hysteria against GMOs. As practical application was delayed year after year we had to conduct socio-economic ex-ante studies for supportive data on the putative impact of Golden Rice. We realized that with the hostile atmosphere towards transgenic plants, created in developing countries by activists from Europe, we had to prepare the organization of social marketing years ahead of release, not to leave the political scene to the GMO-opposition. And there were many more tasks for which a university professor in plant biology was not at all prepared and qualified.

We had a chance to progress through all these problems only because we could, with the help of Dr. Adrian Dubock (Syngenta) establish a public-private-partnership with industry on the basis of transferring the rights for commercial exploitation of our invention to the private partner, for support of the humanitarian project in return. We established a Golden Rice Humanitarian Board with expertise in all the different areas of necessary activity to have expert advice for strategic decisions. We created a Humanitarian Golden Rice Network of 16 public rice research institutes in Southeast Asian countries. And we received support for the appointment of a project manager and a network coordinator, an important addition because our project increasingly also required managerial capacity. While focusing on product development and deregulation of agronomically improved and optimised national Golden Rice varieties to be handed out to the farmers in India, Bangladesh, and The Philippines hopefully by the year 2012, and followed by China, Vietnam, and Indonesia by 2014, the scientific progress has been channelled into a large international programme on 'bio-fortification' (improvement of the micro-nutrient content of basic crop plants on a genetic basis), mainly funded by the Gates Foundation. In this project vitamin A is complemented by the addition of 'high iron', 'high zinc', and 'high quality protein', to work against the other big micro-nutrient deficiencies. The final task is to combine all these novel quality traits in a single rice variety. To extend this help to other poor societies, not dependent upon rice, the same approach is taken with sorghum, cassava, and banana. Bio-fortification is considered by the International Food Policy Research Institute, Washington, the only sustainable solution to the problem because it does not require recurrent financial support once the bio-fortified varieties have been introduced.

All these projects would progress smoothly and rapidly if 'extreme precautionary regulation' of GM (genetically modified) plants were not in place. This process has begun, with no scientific justification, but in response to massive pressure from activists. It is established around the world and it is illegal not to follow these rules and regulations when dealing with plants derived from genetic engineering. The consequence of this situation for Golden Rice is a delay of the use of this technology for a minimum of 7 years, and additional costs of ca. US$ 20 million. The following list gives a few examples of requirements and timeframes, and their scientific justification.

Deletion of selectable marker: unjustified	2 years
Screening for streamlined integration: unjustified	2 years
Screening for regulatory clean events: unjustified	2 years
Protection against liability problems: justified	1 year
Trans-boundary movement of seeds: unjustified	2 years
Obligatory sequence greenhouse-field: unjustified	1 year
Permission for working in the field: unjustified	2 years
Requirement for one-event selection: unjustified	2 years
Experiments for the regulatory dosier: only partly justified	4 years
Deregulation procedure: only partly justified	1 year

This delay is extremely unfortunate because it costs the lives of many children. We know from a state-of-the-art socio-economic ex-ante study for India (A. Stein *et al.*, Nature Biotechnology 2006) that Golden Rice, once established and with strong governmental support, could save in India alone up to 40,000 lives per year, not to mention all those lives lost in the other target countries such as Bangladesh, The Philippines, Vietnam, Indonesia, Nepal, to mention only those for which specific variety development is in progress. The prize for GMO-regulation and the single example of Golden Rice is in the hundreds of thousands of lives. There is no justification to use any hypothetical risk (no concrete risk has been attributed to Golden Rice) to justify the deaths of so many children. The situation becomes far worse when we include all the other possible cases of helpful transgenic plants developed by public institutions in developing countries, which are blocked by regulation. And it comes to extremes when projecting a few years ahead and considering all the upcoming 'bio-fortified' plants, which could reduce not only vitamin A-malnutrition, but also iron-, zinc-, and protein malnutrition, which combined are responsible for the daily death toll of 24,000. There is, therefore, a moral obligation to revise GMO-regulation, and accept that, since the onset of the work with this technology, we have accumulated a wealth of knowledge and experience, which all support the view of the US Academy of Sciences from 12 years ago: GMO technology has no specific, technology-immanent risks, and regulation should be based on novel traits and not on the technology used to acquire the trait.



Potential Impact and Cost-effectiveness of Golden Rice in India.

A.J. Stein, H.P.S.Sachdev, M.Qaim, Nature

Annual burden of vitamin A-deficiency in India:

Lives lost: 71 600

DALYs lost: 2 328 000

Potential annual impact of Golden Rice:

Lives saved: 5 500 to 39 700

DALYs gained: 204 000 to 1 382 000

Cost-effectiveness per DALY‘s saved

WHO standard: $ 620 - 1‘860

World Bank benchmark: $ 200

Supplementation costs: $ 134 - 599

Golden Rice: $ 3 – 19

I am deeply honoured by the invitation to join your most prestigious Academy and I am looking forward to future participation – and hopefully also contribution.

THEODOR W. HÄNSCH

Born on Oct. 30, 1941 at Heidelberg, Germany, Theodor W. Hänsch received his doctor's degree from the University of Heidelberg, Germany, in 1969. In 1970, he came to Stanford University as a postdoctoral fellow, where he was appointed Associate Professor of Physics in 1972. From 1975 to 1986 he held a tenured appointment as a Full Professor in the Department of Physics at Stanford University. In 1986, he returned to his native Germany to become Director at the Max-Planck-Institut für Quantenoptik in Garching and Professor of Physics at the Ludwig-Maximilians-Universität in Munich. Since 1993, he has held a part-time appointment as Professor of Physics at the University of Florence, Italy.

Prof. Hänsch has authored and co-authored more than 450 papers, focusing on coherent nonlinear interactions between light and matter. He is widely known for his seminal contributions in the field of laser spectroscopy. His early work includes the first narrowband tunable dye laser, the invention of commonly used techniques of Doppler-free laser spectroscopy, and the first proposal for laser cooling of atomic gases. Since the early 1970s, Hänsch has pursued precision spectroscopy of the simple hydrogen atom, which permits unique confrontations between experiment and fundamental theory. This work has yielded accurate values of the Rydberg constant, the Lamb shift of the hydrogen ground state, and the charge radii of proton and deuteron. More recently, he has pioneered the revolutionary frequency comb technique for measuring the frequency of light with ultrashort pulses. Exploring the quantum physics of cold neutral atoms, Hänsch and his coworkers have realized the first two- and three-dimensional atomic lattices bound by light, they have demonstrated the first atom laser that emits a continuous beam of coherent matter waves, and they have shown how to integrate a quantum laboratory for ultracold atoms on a microfabricated 'atom chip'. With a Bose-Einstein condensate in an optical lattice potential, they have been the first to observe a quantum phase transition between a wave-like superfluid state and a particle-like Mott insulator crystal.

In 2005, Prof Hänsch was awarded the Physics Nobel Prize jointly with Roy Glauber and John L. Hall *'for his contributions to the development of laser-based precision spectroscopy, including the optical frequency comb technique'.*

THE PIUS XI MEDAL AWARD

Ashoke Sen

Ashoke Sen is an Indian theoretical physicist born in Calcutta in 1956. His main area of interest is string theory, a theory that tries to give a unified description of all matter and the forces between them, based on the postulate that the elementary constituents of matter are tiny, one-dimensional (string-like) objects instead of point particles. He co-discovered S-duality and has proposed a successful explanation of open string tachyon condensation, as well as researching black hole entropy. He has also co-written many papers on string field theory. Sen was awarded the ICTP Prize in 1989. He is currently active at the Harish-Chandra Research Institute (HRI). He is married to Dr. Sumathi Rao, a condensed matter physicist at HRI.

Sen received his PhD from the State University of New York at Stony Brook. During his early career, he worked as a research scientist at Fermilab and the Stanford Linear Accelerator Center (SLAC). Later he joined the Indian Tata Institute of Fundamental Research (TIFR) before finally moving to the HRI. In 1998 he was made a Fellow of the Royal Society.

Research Summary

I have been working exclusively on the subject of string theory since 1985. My first major project in this field involved studying the relationship between the two dimensional σ-models describing string propagation in a given background field, and the space-time properties of these background fields. My main contribution during this project was to establish the relation between classical equations of motion of massless fields in string theory and conformal invariance of the two dimensional sigma model describing string propagation in background of these massless fields. Working along this line I also showed that, in order to get a string compactification that preserves $N=1$ spacetime supersymmetry, the corresponding two dimensional σ-model has (2,0) world-sheet supersymmetry. This provided a way of looking for spacetime supersymmetric vacua of string theory.

My second major project in string theory involved developing a method for generating new classical solutions of string theory from a known classical solution, when the original solution is independent of some of the space-time coordinates. Later, I used this method to generate the most general electrically charged rotating black hole solution in four dimensional heterotic string theory.

My third major project has been in the subject of string dualities. Most of the initial development in the subject of string theory was based on perturbation theory, and there was no method known for studying non-perturbative effects in string theory. In 1992 I presented evidence that a specific string theory, obtained by compactifying heterotic string theory on a six dimensional torus, has a symmetry that relates the strong coupling behaviour of this theory to its weak coupling behaviour. This conjectured symmetry can be used to understand non-perturbative behaviour of string theory. Although initially the evidence for this conjecture was not very strong, in 1994 I showed that this conjecture leads to some precise prediction about the properties of some abstract manifolds (moduli spaces of multi-monopole solutions), and explicitly verified some of these predictions.

Soon after this paper Hull and Townsend – and later Witten – conjectured the existence of many other new duality symmetries, which may sometime relate even different string theories. One of these conjectures stated that the type IIA string theory, compactified on a complicated four dimensional manifold, known as K3, is related to the heterotic string theory compactified on a four dimensional torus. I found non-trivial evidence for this conjecture by showing that the fundamental heterotic string arises as a soliton solution of the type IIA string theory on K3, and that the fundamental type IIA string arises as a soliton solution of the heterotic string theory compactified on T^4.

My fourth major project involves an attempt to understand the Bekenstein-Hawking entropy of black holes from counting the microscopic states in string theory. String theory contains black hole solutions which carry the same quantum numbers as elementary string states. Thus it is natural to ask if the degeneracy of black hole states, as counted by the Bekenstein-Hawking entropy, agrees with the degeneracy of elementary string states. If true, this will indicate that there is no distinction between the black holes and elementary string states, and at the same time, this would provide a statistical interpretation of Bekenstein-Hawking entropy from the counting of microscopic states. The main obstacle to this calculation had been that the degeneracy of elementary string states is calculable only in the weak coupling limit, whereas these states become black holes only for sufficiently large coupling when the gravitational effects are appreciable. I circumvented this problem by looking at the states which preserve part of the spacetime supersymmetry (also known as BPS states), since it is known that for such states the degeneracy remains unchanged as we go from the strong to the weak coupling. Comparison of the black hole entropy according to (a stringy modification of) the Bekenstein-Hawking prescription, and the logarithm of the degeneracy of the elementary string states, showed an exact agreement between the two sides as functions of three independent parameters, – the mass and charge of the black hole, and the string coupling constant – upto an overall multiplicative numerical coefficient which could not be calculated explicitly. (This factor has been calculated recently by Dabholkar). Later similar agreement was found by other authors in many other examples, including the numerical factor where it could be calculated.

My fifth major project involves study of non-supersymmetric solitons in string theory. Most of the earlier studies on solitons in string theory have been on supersymmetric (also known as BPS) configurations. In a series of papers in 1998 I showed how stable non-BPS states can also be used to test various duality conjectures. During this study, I also found a novel construction of non-BPS states in terms of kink solution involving the tachyon field on a brane anti-brane pair. This study led to a series of conjectures about tachyon potential on the brane-antibrane system and non-BPS D-branes in superstring theory, as well as on D-branes of bosonic string theory. Later, in various collaborations with Zwiebach, Berkovits, and Moeller I found evidence for these conjectures in string field theory.

Although initial studies of the non-BPS branes focussed on their static properties, in 2002 I found a set of time dependent solutions describing the ‘decay’ of these branes. These are among the few time dependent solutions in string theory whose properties have been studied in detail and have been used extensively to build cosmological models out of string theory. Study of these solutions has also led to a new kind of duality conjecture between open and closed string theories and is currently under intense investigation.

My sixth major project has been on the study of entropy of extremal black holes in the presence of higher derivative terms. In 2005 I showed that in theories of gravity coupled to other matter fields with generally covariant higher derivative corrections, the near horizon field configuration of an extremal black hole is obtained by extremizing an ‘entropy function’. The entropy function is a function of the parameters characterizing the near horizon geometry of the black hole and there is a well defined algorithm for constructing this function from the lagrangian density of the theory. Furthermore the entropy itself is given by the value of the entropy function at its extremum. This led to a proof of the 'attractor mechanism' in a general higher derivative theory of gravity without invoking supersymmetry. In particular the results show that in a generic situation where the entropy function has no flat directions the near horizon field configuration is determined completely by extremizing the entropy function and hence cannot depend on the asymptotic values of the scalar fields of the theory. On the other hand if the entropy function has flat directions then the near horizon field configuration is not completely determined by extremizing the entropy function and could have some dependence on the asymptotic values of the scalar fields. But the entropy is still independent of the asymptotic data. Although initial studies focussed on spherically symmetric black holes, this analysis has now been generalized to black holes carrying angular momentum.

Besides these six major areas, I have also contributed to some of the more technical aspects of this subject that are listed below.

- In conventional SU(5) grand unified theories, the Higgs field belongs to a fundamental representation of SU(5) and it requires a high degree of fine tuning (1 in 10^{15}) to keep its colour triplet component heavy (which is required to avoid rapid proton decay) and at the same time the weak doublet Higgs light (so that it can induce symmetry breaking responsible for the mass of the $W^\pm$ and Z bosons). I showed how in string theory one might be able to get this mass hierarchy naturally, without the need of any fine tuning.

- In 1986, several authors found a new four loop contribution to the β-function in the σ-model describing string propagation on a Calabi-Yau manifold. This led to the possibility that Calabi-Yau manifolds are not valid backgrounds for string compactification as these would not be solutions of the equations of motion. In collaboration with D. Nemeschansky I showed that it is possible to modify the metric on the Calabi-Yau manifold order by order in string perturbation theory so that it continues to remain solutions of the equations of motion, and hence provides a conformally invariant σ-model.

- In 1987, Dine, Seiberg and Witten used low energy effective field theory to argue that in some four dimensional string theories with U(1) gauge symmetry one loop effects can generate a Fayet-Illiopoulos D-

term that can break supersymmetry. In collaboration with J. Atick and L. Dixon I showed how the presence of such a *D*-term can be verified in an explicit one loop string computation for any string compactification. We also found that for most of the known string theories, the generation of the *D*-term does not break supersymmetry, since one can find a new supersymmetric vacuum in the space of field configurations.

– In 1996 C. Vafa proposed a new way of compactifying type IIB theory known as F-theory. These compactifications are not accessible to the standard perturbative analysis, since the coupling constant of the theory becomes large in some regions in the internal space. Nevertheless based on various symmetry arguments Vafa argued that some of these compactifications are dual to more conventional string compactifications. I showed that at least for some of these compactifications, one can take appropriate limits where they reduce to ordinary string compactifications amenable to perturbative techniques, and the dualities proposed by Vafa can be understood in terms of more conventional dualities proposed earlier. This method has been used later to find various other dualities involving *F*-theory, and has also led to the discovery of new string compactifications in the search for duals of *F*-theory compactification. Using the method of this paper I later showed how one can take appropriate limit of a general *F*-theory compactification to map it into an orientifold.

– In 1996, T. Banks, W. Fischler, S. Shenker and L. Susskind proposed a nonperturbative definition of eleven dimensional supergravity theory in terms of quantum mechanics of infinite dimensional matrices. I gave a systematic description of this theory when we compactify some of the eleven dimensions. This unified many of the ad hoc descriptions of this theory given earlier.

– $\mathcal{N}=4$ supersymmetric string theories typically contain a spectrum of dyon states which preserve 1/4 of the supersymmetries of the original theory. In collaboration with Justin David and Dileep Jatkar I computed the exact spectrum of dyons in a class of such string theories and verified the duality invariance of the spectrum.

SCIENTIFIC PAPERS

Session I

CHAOS AND PREDICTIONS IN PHYSICS AND ASTRONOMY

PREDICTIONS IN ASTROPHYSICS AND COSMOLOGY

RUDOLF MURADIAN

Introduction

In 1962 an unexpected paradigm change occurred in particle physics: G. Chew and S. Frautschi discovered that spin and mass of hadrons are not independent quantities. Experimentally observed mesons and baryons appear to lie on nearly linear and parallel Regge trajectories. This has served as a source of inspiration for the present author for suggesting a remarkably simple cosmic Chew-Frautschi type spin-mass plot for astronomical objects – galaxies, stars and planets. Two fundamental points exist on this cosmic Chew-Frautshi plot, connected with Chandrasekhar and Eddington masses with corresponding angular moments, revealed by the author.

A Portion of History

> *The formation of a world starts with a rotatory motion*
> *which Nous (Νους) imparts to a Chaos (Χαος)* .
> Anaxagoras (500-428 BC)

During his studies of Mandelstam representation in nonrelativistic quantum mechanics Tullio Regge introduced the concept of moving poles in the plane of complex angular momentum. Geoffrey Chew and Steven Frautschi [1] transferred the Regge idea to relativistic hadron physics for grouping together hadronic particles with varying mass and spin into single family on Chew-Frautschi plot. This has had a great impact on the development of elementary particle physics, first leading to the Veneziano amplitude, dual resonance model and then to the concept of relativistic string.

Here we will outline an exciting new insight into the origin of cosmic rotation, provided by the application of the Chew-Frautschi paradigm in an astrophysical context [2]. It will become clear that, without invoking quantum-mechanical concepts, the rotation problem in astrophysics cannot be solved.

The problem of the origin of rotation in stars, galaxies and clusters is *l'Enfant terrible* of physics, astrophysics and cosmology.

What is the source of rotation in Mother Nature? This is an important question for understanding the origin, evolution and structure of celestial bodies and their systems. Rotation is a universal phenomenon and in all scales of the universe: from tiny quarks to huge galaxies we observe rotating objects. The universe is characterized by rotation at every scale: asteroids, planets and their moons, stars, interstellar clouds of gas, globular clusters, galaxies and their clusters rotate around central axes, and things orbit around one another in a hierarchical manner (moons around their planet, planets around their star, stars around the center of their galaxy or globular cluster, galaxies around the center of their galaxy cluster). Understanding the universe is impossible without understanding the source of the rotational motion of cosmic bodies. Spin or angular momentum is a conserved quantity: the total amount of rotation in the whole universe must be constant. Rotation cannot occasionally appear or disappear, but is an innate, inborn, primordial and fundamental entity. When and how was the angular momentum acquired by celestial bodies? Can the rotation serve as the *Rosetta stone* of Astrophysics?

Despite the importance of the problem, surprisingly few attempts have been undertaken to understand the nature and origin of the angular momentum of stars and galaxies from the first principles. The earliest explanation of the origin of galaxy rotation was attempted in 1949 by F. Hoyle. He discussed the possibility of creating an angular momentum from the asymmetric gravitational coupling of protogalaxy with the surrounding matter. This mechanism was reinvestigated in greater details by Peebles (1969). Wieszacker (1951) and Gamow (1952) proposed alternative explanations of galaxy rotation due to primordial turbulence and vortices.

Much of what we present here is based on the work performed by the author in the 1970s-1980s in the Byurakan Astrophysical Observatory, Armenia.

Chew-Frautschi Paradigm

In elementary particle physics after the work of G. Chew and S. Frautschi it has become clear that spin J and mass m of hadrons are not independent but are inherently connected by simple relation [1]

$$J(m^2) = J(0) + J'(0)m^2 \tag{1}$$

where spin J is measured in the units of Planck's constant $\hbar = 1.055\times10^{-34} J{\cdot}s$ and slope J' has the value $J'(0)\approx 1/(GeV/c^2)^2\approx 1/m_p{}^2$, with proton mass $m_p=1.673\times10^{-27}kg$.

Many physicists consider relation (1) as a very fundamental physical law, similar to the cardinal law $E=mc^2$. Neglecting $J(0)$ at large m relation (1) can be essentially rewritten in simpler form [1]:

$$J = \hbar\left(\frac{m}{m_p}\right)^2 \tag{2}$$

This formula is well satisfied by experimental data obtained in high energy physics laboratories, as shown in Fig. 1 (see page 239), where a plot of spin J against the mass (squared) m^2 of different hadrons is represented – a celebrated Chew-Frautschi plot. The meson family falls almost perfectly on a nearly linear Regge trajectory. Mathematically this was like two heavy objects attached to the two ends of a rotating string.

An interesting recollection about the discovery of this relation can be found in the recent interview of S. Frautschi [4]:

> Mandelstam pointed us...that the high-spin particles shouldn't be treated as isolated individuals but as parts of families, and you should organize the calculation around the exchange of the whole family – the spin-one member, the spin-three, the spin-five, and so on...
>
> In particular, our treatment of Regge poles had an equal-spacing feature between masses of successively higher spin – actually, the rule was that the spin went as the mass squared, if you followed the family up to higher masses. Nowadays, both of those developments – exponential growth in particle species and the equal spacing of mass squared in the spin family – are viewed as outgrowths of string theory.
>
> So Geoffrey Chew and I had stumbled upon evidence for strings, although we thought we were working on an entirely different problem.

Geoffrey Chew (see Fig. 2) was a charismatic leader of an 'S-matrix approach' to hadron physics, which, among other things, lead indirectly to the early discovery of *string theory*. Together with several collaborators, in the 1960s-1980s he developed a *bootstrap* approach to subatomic particle physics in which all particles are treated 'democratically', no particle being more fundamental that any other. Contrary to the Greek philosopher Democritus, who conjectured that a reality is constructed out of fundamental building blocks called atoms, Chew's doctrine was in route with another ancient Greek philosopher, Anaxagoras, who postulated the concept of '*Everything in Everything*'.[1]

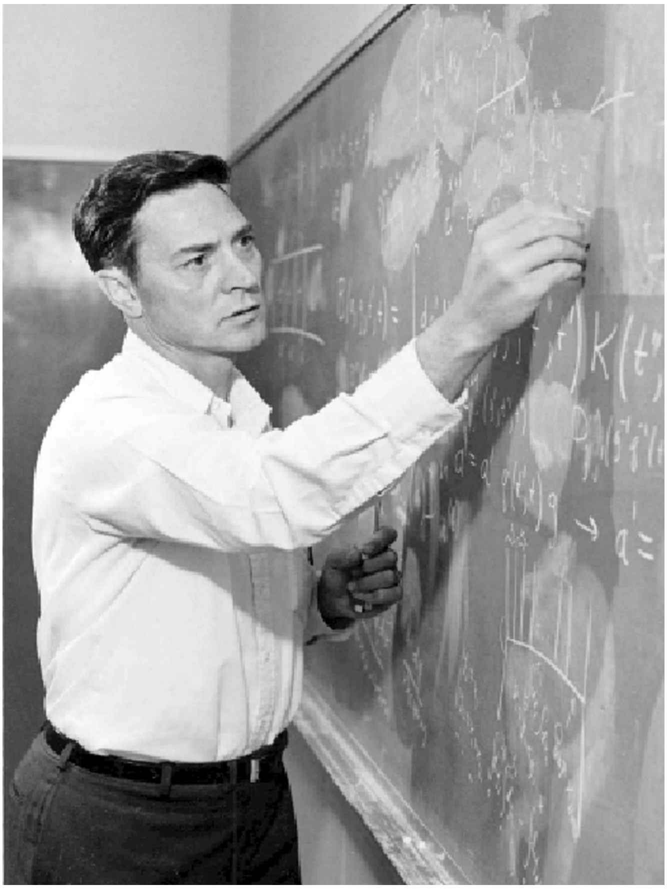


Figure 2. Geoffrey Chew in the 1960s, Lawrence Berkeley National Laboratory, University of California, Berkeley. Courtesy of the Emilio Segré Visual Archives, American Institute of Physics. Taken from [5].

[1] An introduction to the modern string theory can be found in the book [6]. An interesting review, *cum grano salis,* of string theory developments is given in the philosophical and socio-historical essay by Bert Schroer [7].

Bridging Micro and Macro

Dimensional analysis and scaling considerations allowed extending string-like (one-dimensional) formula (1) to multi-dimensional case [2]

$$J = \hbar \left(\frac{m}{m_p} \right)^{1+\frac{1}{n}} \tag{3}$$

Here the exponent *n=1, 2, 3* characterizes the spatial shape of a spinning object.

– The choice *n=1* brings to the previous one-dimensional *rotating string-like* case (2), connected with the usual hadronic Chew-Frautschi plot;

– *n=2* corresponds to the *rotating disk-like* (two-dimensional) configuration

$$J = \hbar \left(\frac{m}{m_p} \right)^{\frac{3}{2}} \tag{4}$$

and describes observational data for galaxies, their clusters and superclusters, and ultimately the universe itself;

– *n=3* corresponds to the *rotating spherical* (three-dimensional) objects

$$J = \hbar \left(\frac{m}{m_p} \right)^{\frac{4}{3}} \tag{5}$$

and well describes spin-mass relation for planets and stars.

The comparison of relations (4) and (5) with observations can be seen in [1] and in other works, cited there. The observational data are fitted well by theoretical formulas maintaining only fundamental constants, without invoking any phenomenological fitting parameters.

Gravitational (Kerr) Angular Momentum

Gravitational or Kerr angular momentum $J_{Kerr} = Gm^2/c$ is a maximal angular momentum of rotating black hole with mass *m*. Here *c* is the speed of light and *G* the gravitational constant. Using Planck mass $m_{Pl} = \sqrt{\hbar c/G}$, where $\hbar$ is Planck's constant, it is possible to rewrite Kerr angular momentum in *string-like* form

$$J_{Kerr} = \hbar \left(\frac{m}{m_{Pl}} \right)^2 \tag{6}$$

This formula formally resembles relation (2) after substitution of proton mass with Planck's one $m_p \rightarrow m_{Planck}$. The huge difference is only in the slope of the trajectory, which is obvious from identity

$$J = \hbar \left(\frac{m}{m_p} \right)^2 \equiv \frac{\hbar c}{G m_p{}^2} \frac{G m^2}{c} = \frac{\hbar c}{G m_p{}^2} J_{Kerr} \qquad (7)$$

The dimensionless combination $\hbar c / G m_p{}^2 = 1.7 \times 10^{38}$ expose the difference of slopes amid hadronic and gravitational strings.

Two Important Points

Equating Kerr momentum with (4) and (5) gives two equations, from which coordinates of intersection points can be deduced (see details in [1]):

$$Chandrasekhar\ mass = m_p \left(\frac{\hbar c}{G m_p{}^2} \right)^{\frac{3}{2}} \qquad spin = \hbar \left(\frac{\hbar c}{G m_p{}^2} \right)^2 \qquad (8)$$

$$Eddington\ mass = m_p \left(\frac{\hbar c}{G m_p{}^2} \right)^2 \qquad spin = \hbar \left(\frac{\hbar c}{G m_p{}^2} \right)^3 \qquad (9)$$

The Chandrasekhar and Eddington masses are one of the most remarkable numbers in all of physics. The corresponding spin expressions were obtained by the author [2].

It is interesting to note, that that all these relations can be represented in terms of Planck mass as follows

$$Chandrasekhar\ mass = m_{Pl} \left(\frac{m_{Pl}}{m_p} \right)^2 \qquad spin = \hbar \left(\frac{m_{Pl}}{m_p} \right)^3 \qquad (10)$$

$$Eddington\ mass = m_{Pl} \left(\frac{m_{Pl}}{m_p} \right)^3 \qquad spin = \hbar \left(\frac{m_{Pl}}{m_p} \right)^4 \qquad (11)$$

Oort, Ambartsumian, E. Burbige, Hoyle and the Rotation Problem

In April 1970 the Pontifical Academy of Sciences organized a *Study Week on Nuclei of Galaxies* [8] with the participation of the prominent astronomers of the time. Among other things, they debated the talk of Soviet-Armenian astronomer Victor Ambartsumian about the possible origin and formation of galaxies due to the activity of their nuclei. This view is completely different from the classical one, according to which galaxies condensed from primeval nebulae.

The objection to such a possibility was formulated on the basis of the conservation law of angular momentum. Indeed quite small dense objects could not have sufficient angular momentum to feed the whole galaxy. Let us show a small excerpt from this discussion:

Oort: Prof. Ambartsumian spoke about the possibility of making a whole galaxy from nucleus by eruption. There is one great difficulty, which is to get *angular momentum*. Angular momentum is such a characteristic thing everywhere in the universe, especially for spiral galaxies, that to me this forms a very great difficulty.

Ambartsumian: On the *angular momentum* which Prof. Oort mentioned: of course, I also keep this problem in my mind…I agree that what I have said is not the real explanation. The situation is dark but there many possibilities.

Oort: I agree that often quite unexpectedly things which we cannot imagine have turned up.

The possible solution of this old quandary lies completely out of the scope of classical physics and can be obtained only in the framework of quantum theory.

Epilogue

> *The difficulty lies, not in the new ideas,*
> *but in escaping the old ones…*
> (John Maynard Keyns)

We have presented a new, *quantum-mechanical* model for the origin of the angular momentum of celestial bodies. Unlike to the previous classical attempts, our approach gives surprisingly accurate numerical predictions of the angular momentum for all spinning astrophysical objects. This occurs for the first time in the history of astronomy.

The creation of spin (angular momentum) is impossible through any applications of a classical field. Artificial invention (postulation) of torque fields such as shear is a unique way to create spin classically.

Another interesting result from this advance is merely philosophical and bears to witness to the unity and simplicity of Nature in micro and macro scales. An understanding of this cannot be achieved by focusing narrowly on the *classical* side of the subject. Instead an integrated, interdisciplinary, open-minded of *quantum-mechanical* vision of the problem of origin of rotation in astrophysics is necessary.

REFERENCES

1. G. Chew, S. Frautschi, Principle of equivalence for all strongly interacting particles within S-matrix framework, *Phys. Rev. Letters*, 7, 394-397, 1961.
2. R. Muradian, Going from quarks to galaxies: two findings, *Paths of Discovery*, Pontificiae Academiae Scientiarum Acta 18, Plenary Session 5-8 November 2004, p. 34.
3. P. Desgrolard, M. Giffon, E. Martynov, E. Predazzi, Exchange-degenerate Regge trajectories, http://arxiv.org/abs/hep-ph/0006244
4. Steven C. Frautschi, *Interviewed by Shierly K. Cohen,* June, 2003, Archives California Institute of Technology, Pasadena, http://oralhistories.library.caltech.edu/120/01/Frautschi_OHO.pdf
5. D. Kaiser, Nuclear democracy: political engagement, pedagogical reform, and particle physics in postwar America, http://web.mit.edu/dikaiser/www/NucDem.pdf
6. *Quantum Fields and Strings: A Course for Mathematicians,* Cambridge University Press, edited by P. Deligne, P. Etingof, D.S. Freed, L.C. Jeffrey, D. Kazhdan, J.W. Morgan, D.R. Morrison and E. Witten
7. Bert Schroer, String theory and the crisis in particle physics, http://www.math.columbia.edu/~woit/schroer.pdf
8. Ambartsumian, V.A., Introduction, Pontificiae Academiae Scientiarum Scripta Varia, *Proceedings of a Study Week on Nuclei of Galaxies*, held in Rome, April 13-18, 1970, Amsterdam: North Holland, and New York: American Elsevier, 1971, edited by D.J.K. O'Connell, p. 9.

COMPLEXITY AND PREDICTIONS AT THE FUNDAMENTAL LEVEL OF SCIENTIFIC KNOWLEDGE

ANTONINO ZICHICHI[*]

1. THE BASIC POINTS

What is the experimental evidence for *Complexity* to *exist*, and for *predictions* to *exist*?

The experimental evidence for the *existence* of *Complexity is as follows*:

1) The Anderson-Feynman-Beethoven-type phenomena (AFB) i.e. phenomena whose laws and regularities ignore the existence of the Fundamental Laws of Nature from which they originate (see chapter 2);

2) The Sarajevo-type effects, i.e. Unexpected Events of quasi irrelevant magnitude which produce Enormous Consequences (UEEC) (see chapter 3).

The experimental evidence for the *existence* of *predictions* consists of the very many results of reproducible scientific experiments.

For example the measurement of the anomalous magnetic moment, in symbols (g–2), of the electron (e):

$$(g\text{–}2)_e$$

which is theoretically computed at an extraordinary level of precision (few parts in ten billion parts) and is experimentally verified to be correct.

[*] University of Bologna, Italy; INFN (National Institute of Nuclear and Subnuclear Physics), Rome, Italy; Enrico Fermi Centre, Rome, Italy; CERN (European Centre for Nuclear and Subnuclear Research), Geneva, Switzerland; EMFCSC (Ettore Majorana Foundation and Centre for Scientific Culture), Erice, Italy; WFS (World Federation of Scientists), Beijing, Geneva, Moscow, New York.

Could the

$$(g\text{–}2)_e$$

be predicted before the discovery of the Maxwell equations and the existence of Quantum ElectroDynamics (QED)?

Predictions at the *fundamental level of scientific knowledge* depend on *UEEC events*.

For example: it is the discovery of the laws governing electric, magnetic and optical phenomena (all totally unpredicted) which produced the mathematical structure called QED.

The mathematical structure was not invented before the innumerable series of *UEEC events* in electricity, magnetism and optics which allowed Maxwell to express 200 years of experimental discoveries in a set of 4 equations.

The mathematical formalism comes *after* a totally unexpected discovery: an *UEEC event* which no one was able *to predict*.

In the whole of our knowledge predictions exist only in Science.

These predictions are the analytic continuation of what is already known. The *greatest* steps in the *progress of Science* come from totally unpredicted discoveries.

This is the reason why we need to perform experiments, as Galileo Galilei realized, 400 years ago.

Today we have all the mathematics needed to describe the *Superworld* but in order to know if the Superworld exists we need the experimentally reproducible proof of its existence.

2. AFB Phenomena from Beethoven to the Superworld

Let me now mention a few examples of AFB phenomena in Science.

Beethoven and the laws of acoustics.

Beethoven could compose superb masterpieces of music without any knowledge of the laws governing acoustic phenomena. But these masterpieces could not exist if the laws of acoustics were not there.

The living cell and QED.

To study the mechanisms governing a living cell, we do not need to know the laws of electromagnetic phenomena whose advanced formulation is called Quantum ElectroDynamic, QED.

All mechanisms needed for life are examples of purely electromagnetic processes. If QED were not there Life could not exist.

Nuclear Physics and QCD.

Proton and neutron interactions appear as if a fundamental force of nature is at work: the nuclear force, with its rules and its regularities.

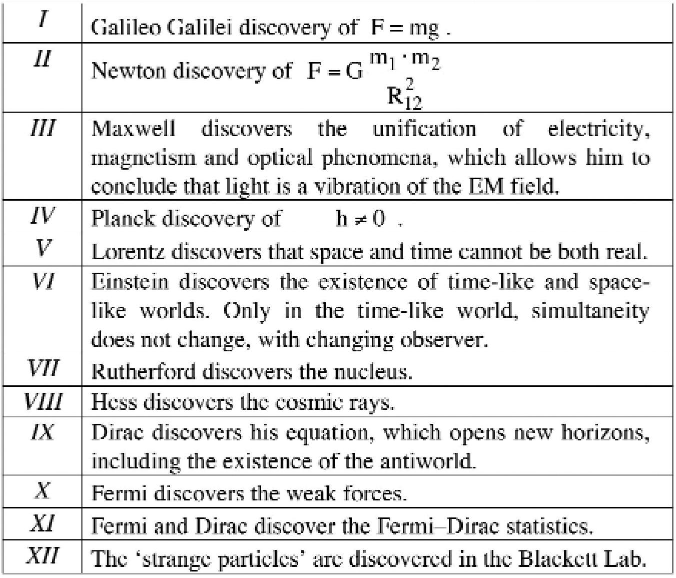
These interactions ignore that protons and neutrons are made with quarks and gluons.

Nuclear physics does not appear to care about the existence of QCD, although all phenomena occurring in nuclear physics have their roots in the interactions of quarks and gluons.

In other words, protons and neutrons behave like Beethoven: they interact and build up nuclear physics without 'knowing' the laws governing QCD. The most recent example of an Anderson-Feynman-Beethoven-type phenomenon: *the world could not care less about the existence of the Superworld*.

3. UEEC Events, from Galilei up to the Present Day

In figure 1 there is a sequence of UEEC events from Galilei to Fermi-Dirac and the 'strange particles'. In figures 2, 3, 4 from Fermi-Dirac to the construction of the Standard Model and in figure 5 a synthesis of the UEEC events in what we now call the Standard Model and Beyond (SM&B).



I	Galileo Galilei discovery of $F = mg$.
II	Newton discovery of $F = G \frac{m_1 \cdot m_2}{R_{12}^2}$
III	Maxwell discovers the unification of electricity, magnetism and optical phenomena, which allows him to conclude that light is a vibration of the EM field.
IV *V*	Planck discovery of $h \neq 0$. Lorentz discovers that space and time cannot be both real.
VI *VII*	Einstein discovers the existence of time-like and space-like worlds. Only in the time-like world, simultaneity does not change, with changing observer. Rutherford discovers the nucleus.
VIII	Hess discovers the cosmic rays.
IX	Dirac discovers his equation, which opens new horizons, including the existence of the antiworld.
X	Fermi discovers the weak forces.
XI	Fermi and Dirac discover the Fermi–Dirac statistics.
XII	The 'strange particles' are discovered in the Blackett Lab.

Figure 1. 'UEEC'. Totally Unexpected Discoveries. From Galilei to Fermi-Dirac and the 'Strange' Particles

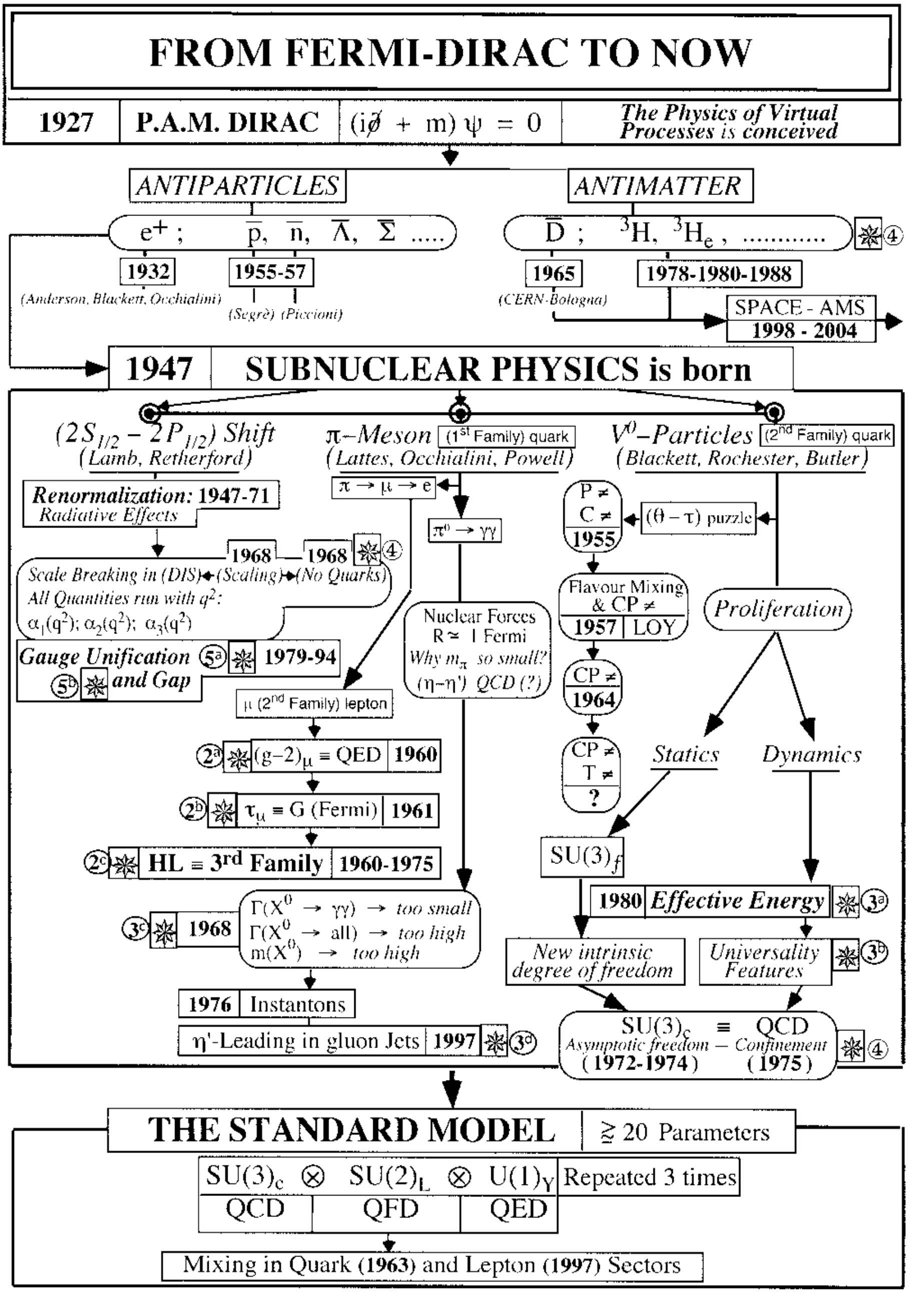

FROM FERMI-DIRAC TO NOW
1927
P.A.M. DIRAC
(i∂̸ + m) ψ = 0
The Physics of Virtual Processes is conceived
ANTIPARTICLES
ANTIMATTER
e+ ; p̄, n̄, Λ̄, Σ̄ .....
D̄ ; ³H̄, ³H̄e , ...
④
1932
(Anderson, Blackett, Occhialini)
1955-57
(Segrè) (Piccioni)
1965
(CERN-Bologna)
1978-1980-1988
SPACE - AMS
1998 - 2004
1947
SUBNUCLEAR PHYSICS is born
(2S1/2 – 2P1/2) Shift
(Lamb, Retherford)
π–Meson
(1st Family) quark
(Lattes, Occhialini, Powell)
V0–Particles
(2nd Family) quark
(Blackett, Rochester, Butler)
Renormalization: 1947-71
Radiative Effects
π → μ → e
π0 → γγ
P ≠
C ≠
1955
(θ – τ) puzzle
1968
1968
④
Scale Breaking in (DIS)↔(Scaling)↔(No Quarks)
All Quantities run with q²:
α1(q²); α2(q²); α3(q²)
Gauge Unification 5a 1979-94
5b and Gap
Nuclear Forces
R ≃ 1 Fermi
Why mπ so small?
(η–η') QCD (?)
Flavour Mixing
& CP ≠
1957 LOY
Proliferation
CP ≠
1964
μ (2nd Family) lepton
2a (g–2)μ ≡ QED 1960
2b τμ ≡ G (Fermi) 1961
2c HL ≡ 3rd Family 1960-1975
CP ≠
T ≠
?
Statics
Dynamics
SU(3)f
3c 1968
Γ(X0 → γγ) → too small
Γ(X0 → all) → too high
m(X0) → too high
1980 Effective Energy 3a
New intrinsic degree of freedom
Universality Features 3b
1976 Instantons
η'-Leading in gluon Jets 1997 3d
SU(3)c ≡ QCD
Asymptotic freedom – Confinement
(1972-1974) (1975)
④
THE STANDARD MODEL
≳ 20 Parameters
SU(3)c ⊗ SU(2)L ⊗ U(1)Y
Repeated 3 times
QCD
QFD
QED
Mixing in Quark (1963) and Lepton (1997) Sectors

Figure 2.

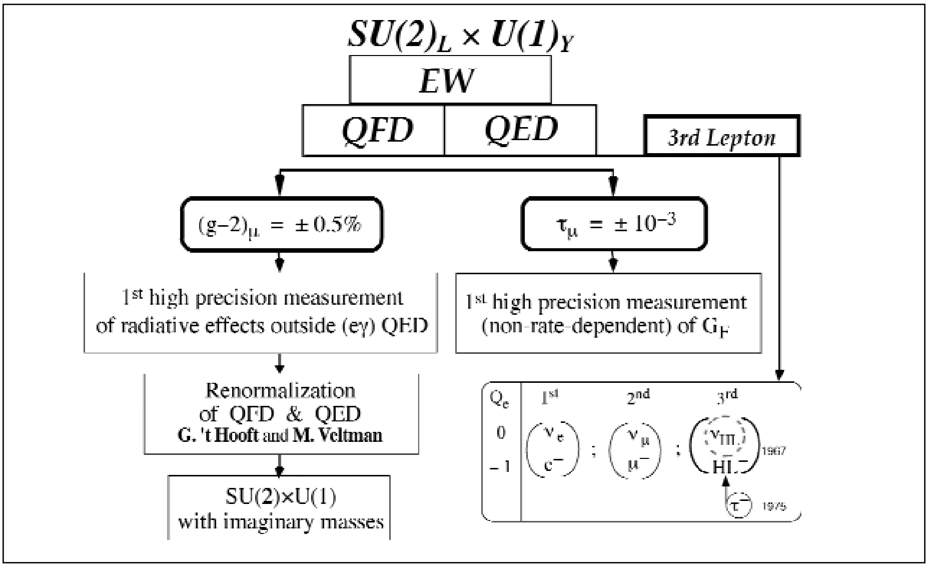

SU(2)L × U(1)Y
EW
QFD
QED
3rd Lepton
(g–2)μ = ±0.5%
τμ = ± 10−3
1st high precision measurement of radiative effects outside (eγ) QED
1st high precision measurement (non-rate-dependent) of GF
Renormalization of QFD & QED
G. 't Hooft and M. Veltman
SU(2)×U(1) with imaginary masses
Qe
0
−1
1st
2nd
3rd
νe e−
νμ μ−
νHL HL
1967
τ−
1975

Figure 3. Details from figure 2, concerning $SU(2)_L$ and $U(1)_Y$.

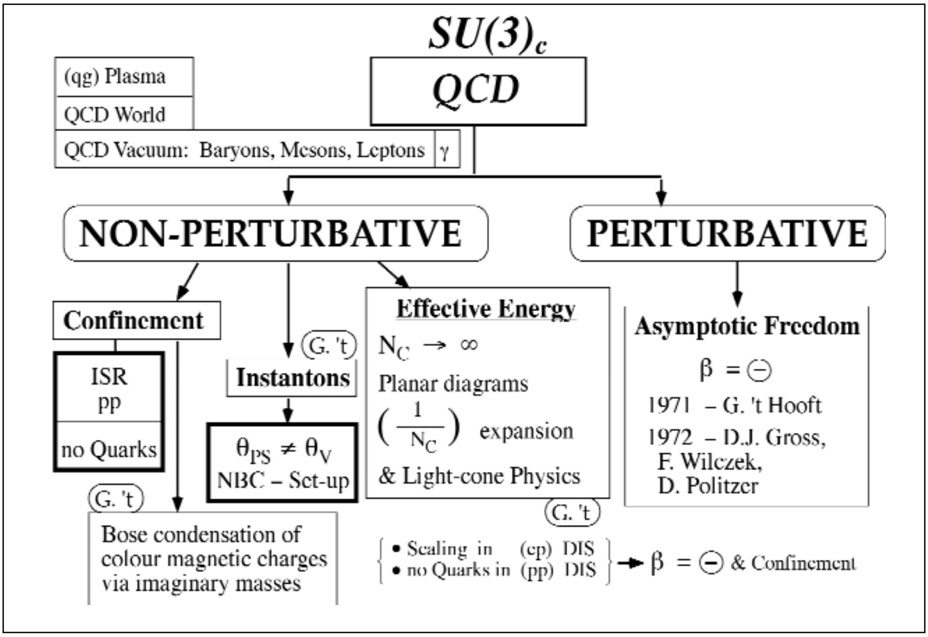
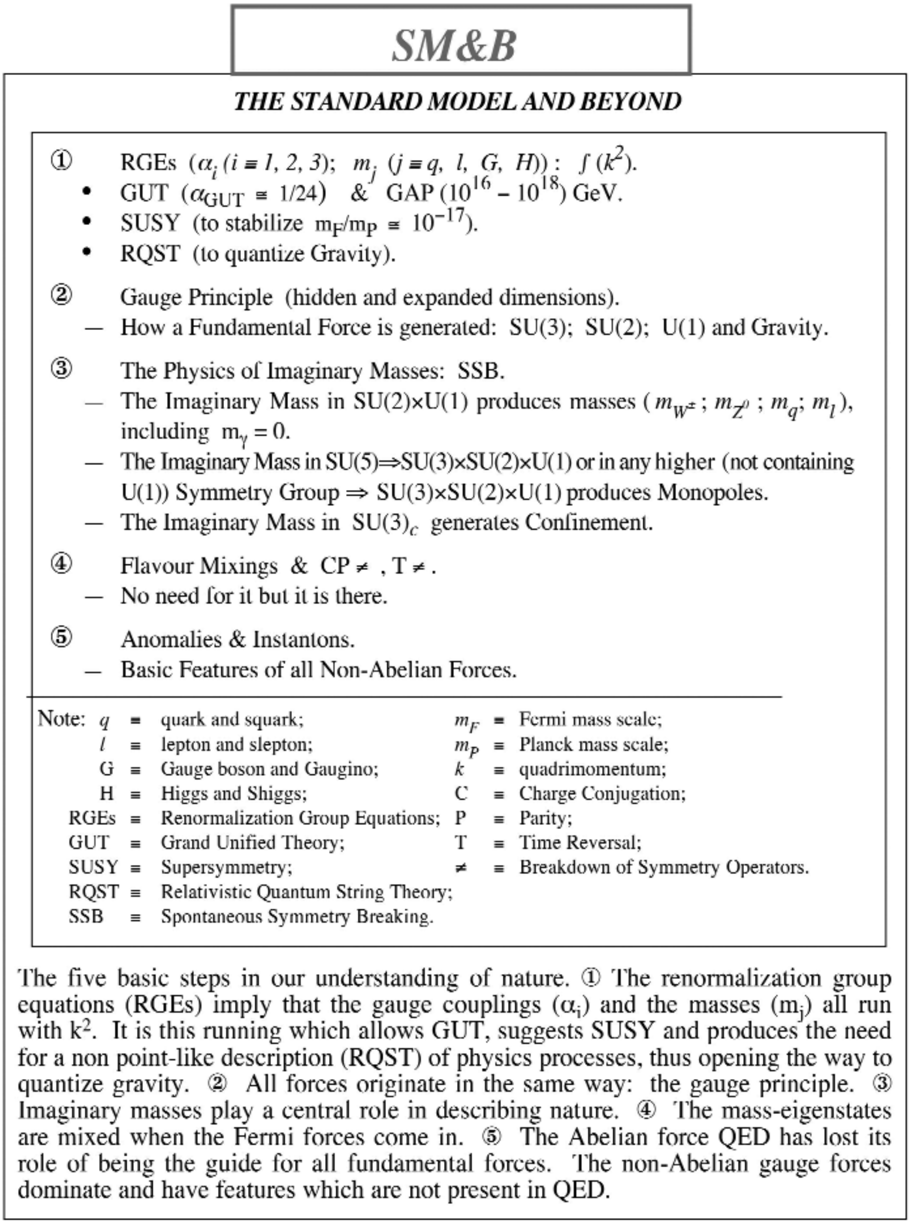


Figure 4. Details from figure 2, concerning $SU(3)_c$.



SM&B

THE STANDARD MODEL AND BEYOND

① RGEs $(\alpha_i\ (i \equiv 1, 2, 3);\ m_j\ (j \equiv q,\ l,\ G,\ H)) :\ f\,(k^2)$.

- GUT $(\alpha_{\rm GUT} \equiv 1/24)$ & GAP $(10^{16} - 10^{18})$ GeV.
- SUSY (to stabilize $m_F/m_P \equiv 10^{-17}$).
- RQST (to quantize Gravity).

② Gauge Principle (hidden and expanded dimensions).

— How a Fundamental Force is generated: SU(3); SU(2); U(1) and Gravity.

③ The Physics of Imaginary Masses: SSB.

— The Imaginary Mass in SU(2)×U(1) produces masses ($m_{W^\pm}$; m_{Z^0}; m_q; m_l), including $m_\gamma = 0$.

— The Imaginary Mass in SU(5)⇒SU(3)×SU(2)×U(1) or in any higher (not containing U(1)) Symmetry Group ⇒ SU(3)×SU(2)×U(1) produces Monopoles.

— The Imaginary Mass in $\mathrm{SU(3)}_c$ generates Confinement.

④ Flavour Mixings & CP ≠ , T ≠ .

— No need for it but it is there.

⑤ Anomalies & Instantons.

— Basic Features of all Non-Abelian Forces.

Note:

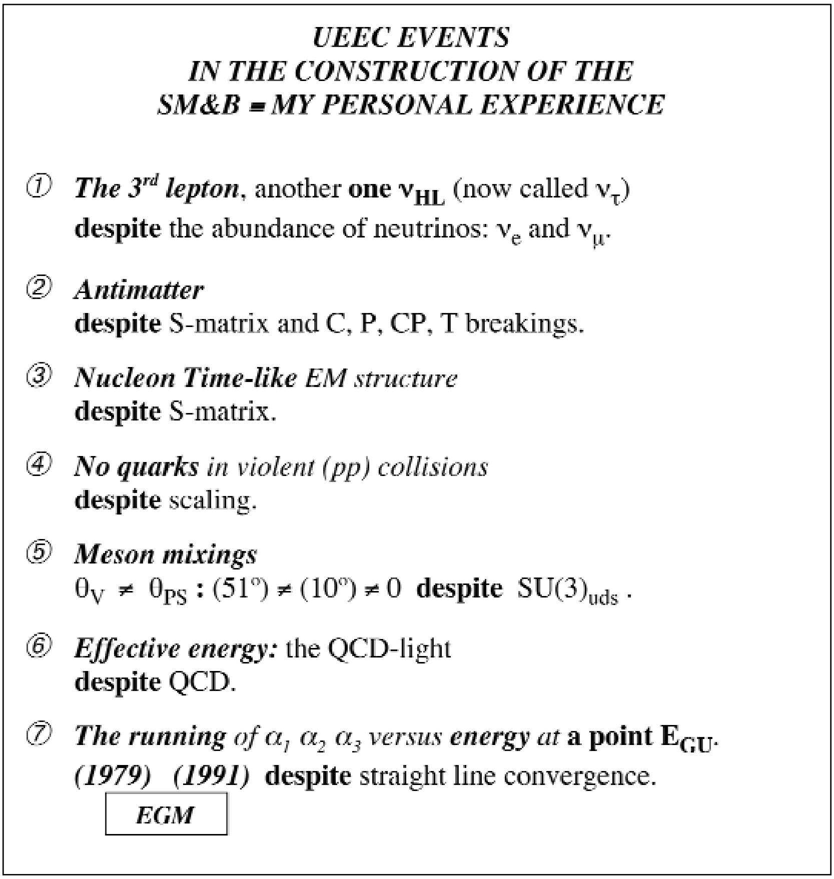
q	≡ quark and squark;	m_F	≡ Fermi mass scale;
l	≡ lepton and slepton;	m_P	≡ Planck mass scale;
G	≡ Gauge boson and Gaugino;	k	≡ quadrimomentum;
H	≡ Higgs and Shiggs;	C	≡ Charge Conjugation;
RGEs	≡ Renormalization Group Equations;	P	≡ Parity;
GUT	≡ Grand Unified Theory;	T	≡ Time Reversal;
SUSY	≡ Supersymmetry;	≠	≡ Breakdown of Symmetry Operators.
RQST	≡ Relativistic Quantum String Theory;		
SSB	≡ Spontaneous Symmetry Breaking.		

The five basic steps in our understanding of nature. ① The renormalization group equations (RGEs) imply that the gauge couplings (α_i) and the masses (m_j) all run with k^2. It is this running which allows GUT, suggests SUSY and produces the need for a non point-like description (RQST) of physics processes, thus opening the way to quantize gravity. ② All forces originate in the same way: the gauge principle. ③ Imaginary masses play a central role in describing nature. ④ The mass-eigenstates are mixed when the Fermi forces come in. ⑤ The Abelian force QED has lost its role of being the guide for all fundamental forces. The non-Abelian gauge forces dominate and have features which are not present in QED.

Figure 5.

A few cases (seven) where I have been directly involved are summarised in figure 6.

Each UEEC event is coupled with a *despite,* in order to emphasize the reason why the event is unexpected.



UEEC EVENTS
IN THE CONSTRUCTION OF THE
SM&B ≡ MY PERSONAL EXPERIENCE

① ***The 3rd lepton***, another **one** $\nu_{\mathbf{HL}}$ (now called ν_τ)
despite the abundance of neutrinos: ν_e and ν_μ.

② ***Antimatter***
despite S-matrix and C, P, CP, T breakings.

③ ***Nucleon Time-like*** *EM structure*
despite S-matrix.

④ ***No quarks*** *in violent (pp) collisions*
despite scaling.

⑤ ***Meson mixings***
$\theta_V \neq \theta_{PS}$ **:** $(51°) \neq (10°) \neq 0$ **despite** $SU(3)_{uds}$.

⑥ ***Effective energy:*** the QCD-light
despite QCD.

⑦ ***The running*** *of* $\alpha_1\ \alpha_2\ \alpha_3$ *versus* ***energy*** *at* **a point** $\mathbf{E_{GU}}$.
(1979) (1991) **despite** straight line convergence.
EGM

Figure 6.

The SM&B is the greatest synthesis of all time in the study of the fundamental phenomena governing the Universe in all its structures.

The basic achievements of the SM&B have been obtained via UEEC events; moreover the SM&B could not care less about the existence of Platonic Simplicity. An example is shown in figure 7 where the straight line (small dots) would be the Platonic simple solution towards the Unification of all Fundamental Forces. But the effective unification is expected to be along the sequence of points (the big ones) calculated using the Renormalization Group Equations (RGEs).

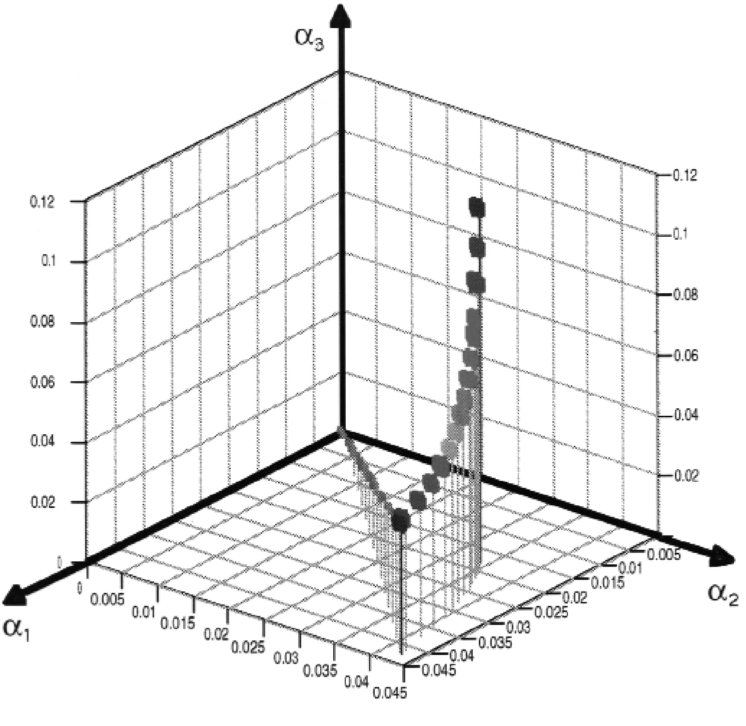


Figure 7. The points have a sequence of 100 GeV in energy. The last point where the 'ideal' platonic straight line intercepts the theoretical prediction is at the energy of the Grand Unification. This corresponds to $E_{GU}=10^{16.2}$ GeV. Other detailed information on the theoretical inputs: the number of fermionic families, N_F, is 3; the number of Higgs particles, N_H, is 2. The input values of the gauge couplings at the Z^0-mass is $\alpha_3(M_Z)=0.118\pm0.008$; the other input is the ratio of weak and electromagnetic couplings also measured at the Z^0-mass value: $\sin^2\Theta_W(M_Z)=0.2334\pm0.0008$.

Platonic Simplicity is violated at every corner in the process of construction of the SM [1].

These violations are the proof that Complexity exists at the fundamental level of scientific knowledge where we have proved that AFB phenomena and UEEC events are present.

The conclusion is that Complexity exists at the elementary level. In fact, starting from Platonic Simplicity, the SM&B needs a series of 'ad hoc' inputs [1].

4. Seven Definitions of Complexity

People speak of 'Complexity' as a source of new insights in physics, biology, geology, cosmology, social sciences and in all intellectual activities which look at the world through the lens of a standard analysis in terms of either Simplicity or Complexity. But 'Complexity' is ill-defined, as shown by the existence of at least seven definitions of Complexity.

Definition Number 1

Complexity is a property of systems that are somewhere in between a completely random and a completely regular state, often described by a highly non linear set of equations but sometimes not describable by equations at all.

Definition Number 2

Bad ones:

1) Chaos.
2) The need for lengthy calculations.
3) The need for many distinct variables.

Better ones:

4) Unexpected difficulty when attempting to describe something in a precisely formulated theory.
5) What is left over after all systematic approaches failed.

But it could also be that: Complexity is an excuse for sloppy thinking.

Definition Number 3

The Complexity of a theory (problem) is the minimum amount of computer time and storage required to simulate (solve) it to a specified level of precision.

Definition Number 4

If we admit that biological or linguistic evolution, or financial dynamics are complex phenomena, then their typical dynamics is somehow between strong chaos (i.e. positive Lyapunov exponents) and simple orbits (i.e. negative Lyapunov exponents). In other words, Complexity (or at least some form of it) is deeply related to the edge of chaos (i.e. vanishing maximal Lyapunov exponent). Since the edge of chaos appears to be related paradigmatically to an entropy index ‘q’ different from unity, there must be some deep connection between Complexity and generalized entropies such as ‘S_q’.

Definition Number 5

From the mathematical point of view:

- A problem can be polinomial, which means that it is not to hard to predict surprises.
- A problem can be NP or NP-complete, which represent different degrees of difficulty in predicting surprises.
- •• Surprises mean: UEEC event.
- •• That degree of difficulty can be associated with the level of Complexity.

Definition Number 6

A system is ‘complex’ when it is no longer useful to describe it in terms of its fundamental constituents.

Definition Number 7

The simplest definition of Complexity: ‘*Complexity is the opposite of Simplicity*’. This is why we have studied the platonic Standard Model and its extension to the platonic Superworld.

These seven definitions of Complexity must be compared with the whole of our knowledge in order to focus our attention on the key features needed to study our real world.

5. Complexity Exists at all Scales

The Logic of Nature allows the existence of a large variety of structures with their regularities and laws which appear to be independent from the basic constituents of Nature and fundamental laws which govern their interactions.

But, without these laws it would be impossible to have the real world which is in front of us and of which we are part of. A series of complex systems is shown in figure 8.

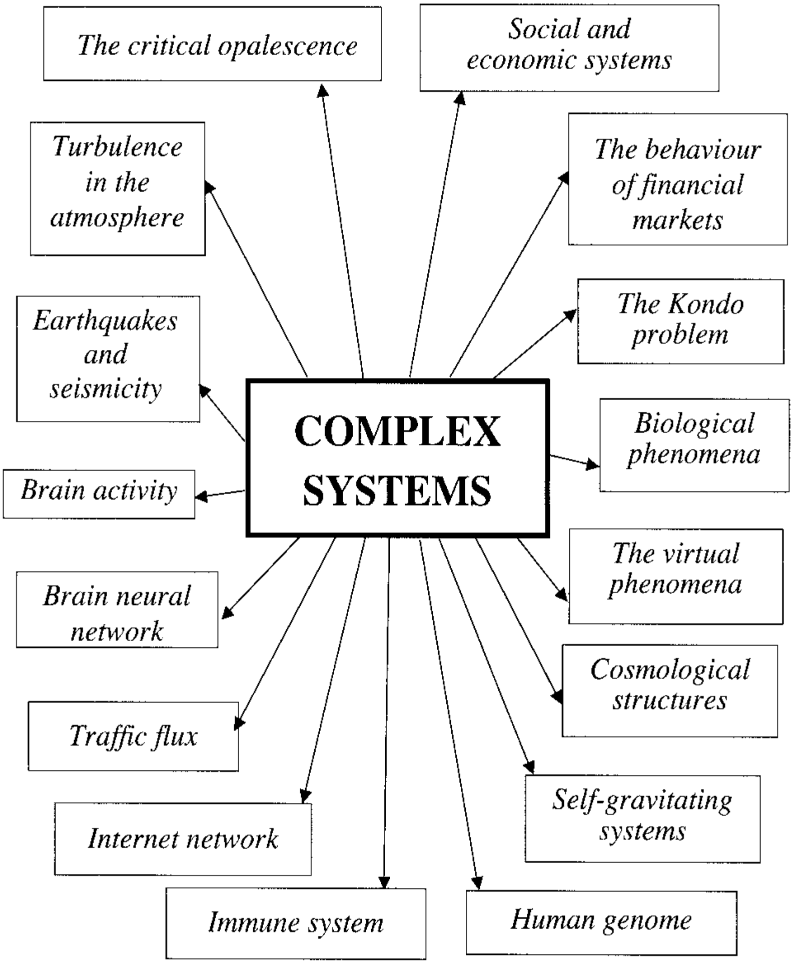


Figure 8.

As you can see, we go from traffic flux, to the internet network, to earthquakes and seismicity, to social and economic systems, to the behaviour of financial markets, to the study of cosmological structures, and so on.

There is no question that nature shows structures which are considered complex on the basis of AFB and UEEC events (as shown in figure 9).

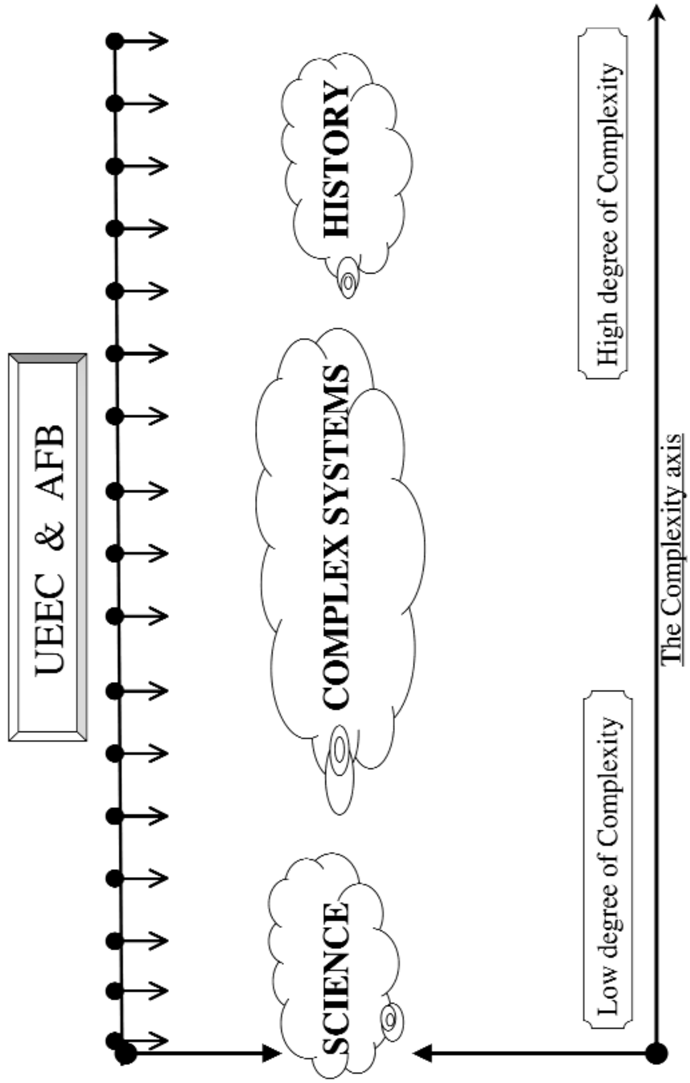


Figure 9.

The only certainty about Complexity is the existence of the experimentally observable effects: UEEC & AFB. These effects exist at all scales, and therefore Complexity exists at all scales, as illustrated in figure 9.

6. Science, from Planck to Complexity

Four centuries of Galilean research work based on Reductionism, i.e. on the identification of the simplest elements in the study of Nature, has allowed us to get the greatest achievement of Science, i.e. the so called *Standard Model* and its extension (SM&B), illustrated before in figure 5.

This extension predicts GUT (the Grand Unification Theory), the existence of the Superworld and the resolution of the quantum-gravity problem via the powerful theoretical structure of RQST (Relativistic Quantum String Theory). All these developments started thirty years ago when a great scientific novelty came; all experimental discoveries obtained with our powerful accelerators were to be considered only matters of extremely low energy.

The scale of energy on which to direct the attention to understand the Logic that rules the world, from the tiniest structures to the galactic ones, had to be shifted at a much higher level: to the mass-energy named after Planck, E_{Planck}, something like seventeen powers of ten above the Fermi scale, E_{Fermi} , that already seemed to be an extremely high level of energy.

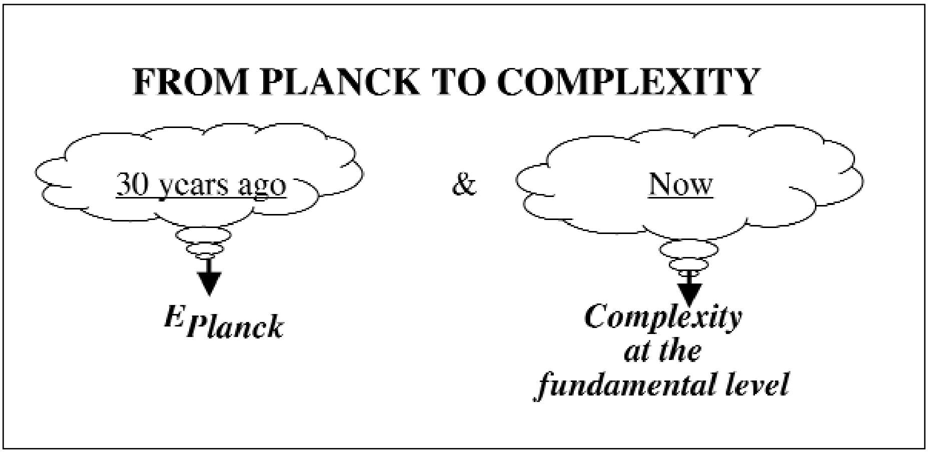


Figure 10.

Now, after thirty years, comes the novelty of our time, illustrated in figure 10: Complexity exists at the fundamental level [1]. In fact, AFB and UEEC events exist at all scales, as reported in chapter 5.

This result is corroborated by the mathematical structure (the only one) that is in a position of describing all that happens at the Planck scale: the Relativistic Quantum String Theory (RQST).

This mathematical structure produces innumerable minima of energy, named *Landscape*.

The theoretical discovery of the *Landscape* (Leonard Susskind) [2], has been followed by another formidable discovery in mathematical physics: the most rigorous model of RQST (Raphael Bousso and Joseph Polchinski) is NP-complete (Michael R. Douglas and Frederik Denef) [3].

This discovery corroborates all that we have put in evidence during the last five years [4-7]: *Complexity exists at the fundamental level* [1].

We do not know what will be the final outcome of String Theory.

What we do know is that: '*The world appears to be complex at every scale.* Therefore *we must expect a continued series of surprises that we cannot easily predict*'.

7. The Two Asymptotic Limits: History and Science

The real world seems characterized by two basic features, which are one on the opposite side of the other: *Simplicity* and *Complexity*.

It is generally accepted that *Simplicity* is the outcome of *Reductionism*, while *Complexity* is the result of *Holism*.

The most celebrated example of *Simplicity* is *Science* while the most celebrated example of *Complexity* is *History*.

Talking about asymptotic limits, the general trend is to consider *History* as the asymptotic limit of *Holism* and of *Complexity*; *Science* as the asymptotic limit of *Reductionism* and of *Simplicity*, as illustrated in figure 11.

The Logic of Nature allows the existence of Science (the asymptotic limit of Simplicity) and of History (the asymptotic limit of Complexity), which share a property, common to both of them.

It is interesting to define Science and History in terms of this property, probably the only one that they share; i.e. Evolution.

– Science is the Evolution of our Basic Understanding of the laws governing the world in its Structure ≡ EBUS.
– History is the Evolution of the World in its Real Life ≡ EWRL.

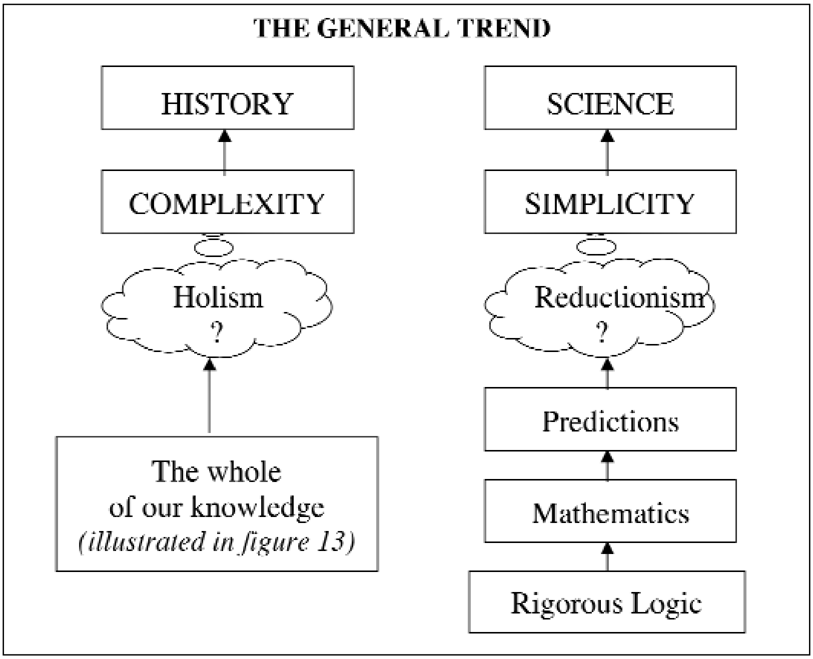


Figure 11.

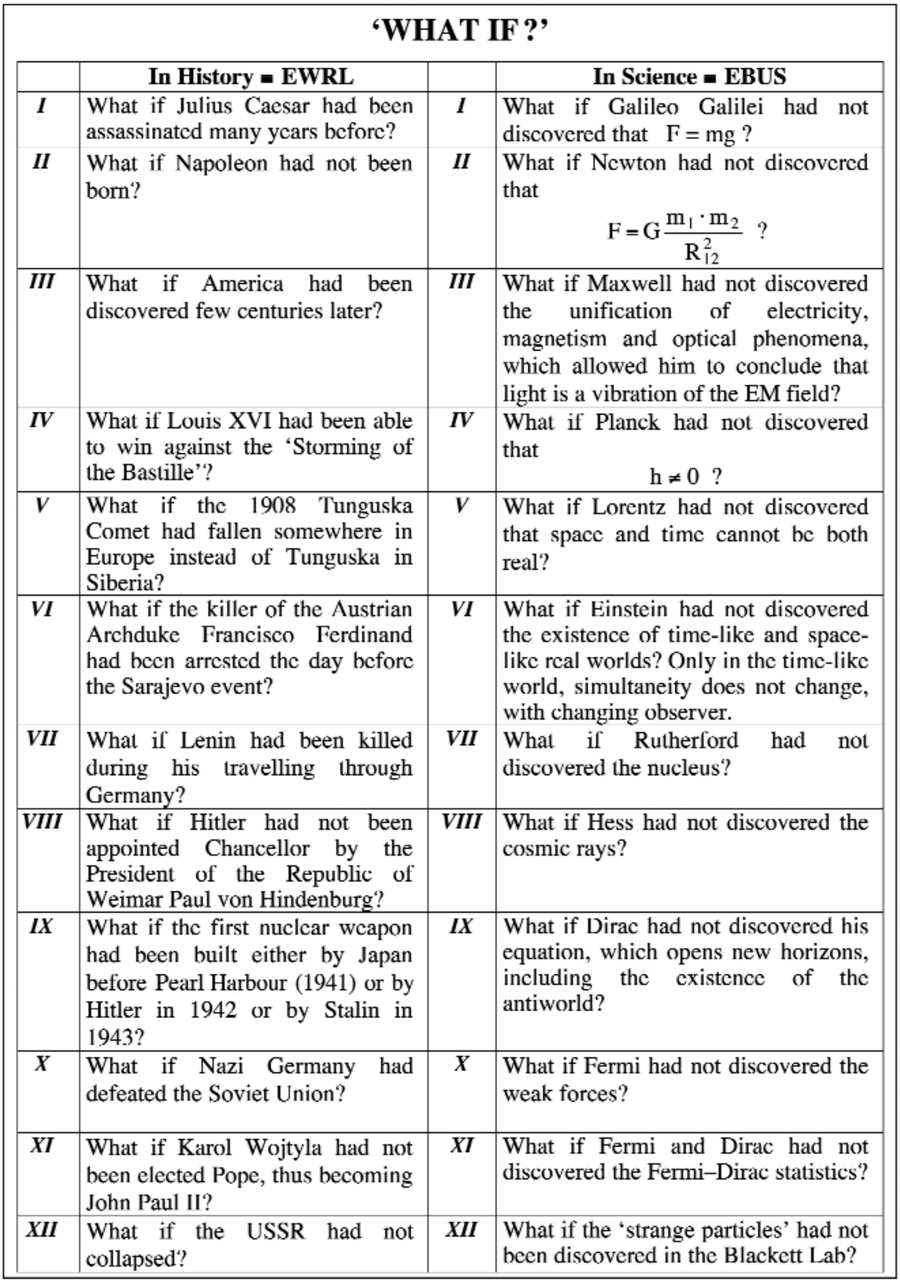
In Table 1 we compare these two supposedly asymptotic limits – History and Science – on the basis of 'What if?'; a condition elaborated by the specialists in what is now known as 'virtual history' [8].

On the basis of 'What if?' these specialists conclude that the world would not be as it is, if one, or few, or any number of 'What if?' had not been as History tells us. This is not the case for Science. The world would have exactly the same laws and regularities, whether Galileo Galilei or somebody else had discovered

$$F = mg,$$

and so on for all the other scientific discoveries.

It is in the consequences of 'What if?' that the two asymptotic limits of Simplicity and Complexity seem to diverge, despite the fact that the sequence of 'What if?' in Science belongs to the 'totally unexpected events' (UEEC) exactly like the others listed in the column of History.



'WHAT IF ?'

	In History ■ EWRL		In Science ■ EBUS
I	What if Julius Caesar had been assassinated many years before?	*I*	What if Galileo Galilei had not discovered that $F = mg$?
II	What if Napoleon had not been born?	*II*	What if Newton had not discovered that $F = G\frac{m_1 \cdot m_2}{R_{12}^2}$?
III	What if America had been discovered few centuries later?	*III*	What if Maxwell had not discovered the unification of electricity, magnetism and optical phenomena, which allowed him to conclude that light is a vibration of the EM field?
IV	What if Louis XVI had been able to win against the 'Storming of the Bastille'?	*IV*	What if Planck had not discovered that $h \neq 0$?
V	What if the 1908 Tunguska Comet had fallen somewhere in Europe instead of Tunguska in Siberia?	*V*	What if Lorentz had not discovered that space and time cannot be both real?
VI	What if the killer of the Austrian Archduke Francisco Ferdinand had been arrested the day before the Sarajevo event?	*VI*	What if Einstein had not discovered the existence of time-like and space-like real worlds? Only in the time-like world, simultaneity does not change, with changing observer.
VII	What if Lenin had been killed during his travelling through Germany?	*VII*	What if Rutherford had not discovered the nucleus?
VIII	What if Hitler had not been appointed Chancellor by the President of the Republic of Weimar Paul von Hindenburg?	*VIII*	What if Hess had not discovered the cosmic rays?
IX	What if the first nuclear weapon had been built either by Japan before Pearl Harbour (1941) or by Hitler in 1942 or by Stalin in 1943?	*IX*	What if Dirac had not discovered his equation, which opens new horizons, including the existence of the antiworld?
X	What if Nazi Germany had defeated the Soviet Union?	*X*	What if Fermi had not discovered the weak forces?
XI	What if Karol Wojtyla had not been elected Pope, thus becoming John Paul II?	*XI*	What if Fermi and Dirac had not discovered the Fermi–Dirac statistics?
XII	What if the USSR had not collapsed?	*XII*	What if the 'strange particles' had not been discovered in the Blackett Lab?

Table 1.

8. Conclusions

We have proved that AFB and UEEC – which are at the origin of Complexity, with its consequences permeating all our existence, from molecular biology to life in all its innumerable forms up to our own, including History – do exist at the fundamental level [4-7] and [1].

It turns out that Complexity in the real world exists, no matter the mass-energy and space-time scales considered.

Therefore the only possible prediction is that:

- *Totally Unexpected Effects* should *show up.*
- *Effects*, which are impossible to be predicted on the basis of *present knowledge*.

We should be prepared with powerful experimental instruments, technologically at the frontier of our knowledge, to discover Totally Unexpected Events in all laboratories, the world over (including CERN in Europe and Gran Sasso in Italy).

The mathematical descriptions, and therefore the predictions come after an UEEC event, never before.

Recall:

- The *discoveries in Electricity, Magnetism* and *Optics* (UEEC).
- *Radioactivity* (UEEC).
- The *Cosmic Rays* (UEEC).
- The *Weak Forces* (UEEC).
- The *Strange Particles* (UEEC).
- The *3 Columns* (UEEC).
- The *origin of the Fundamental Forces* (UEEC).

The present status of Science is reported in figure 12.

It could be that Science will be mathematically proved to be ‘NP-complete’. This is the big question for the immediate future [9].

It is therefore instructive to see how Science fits in the whole of our knowledge as reported in figure 13.

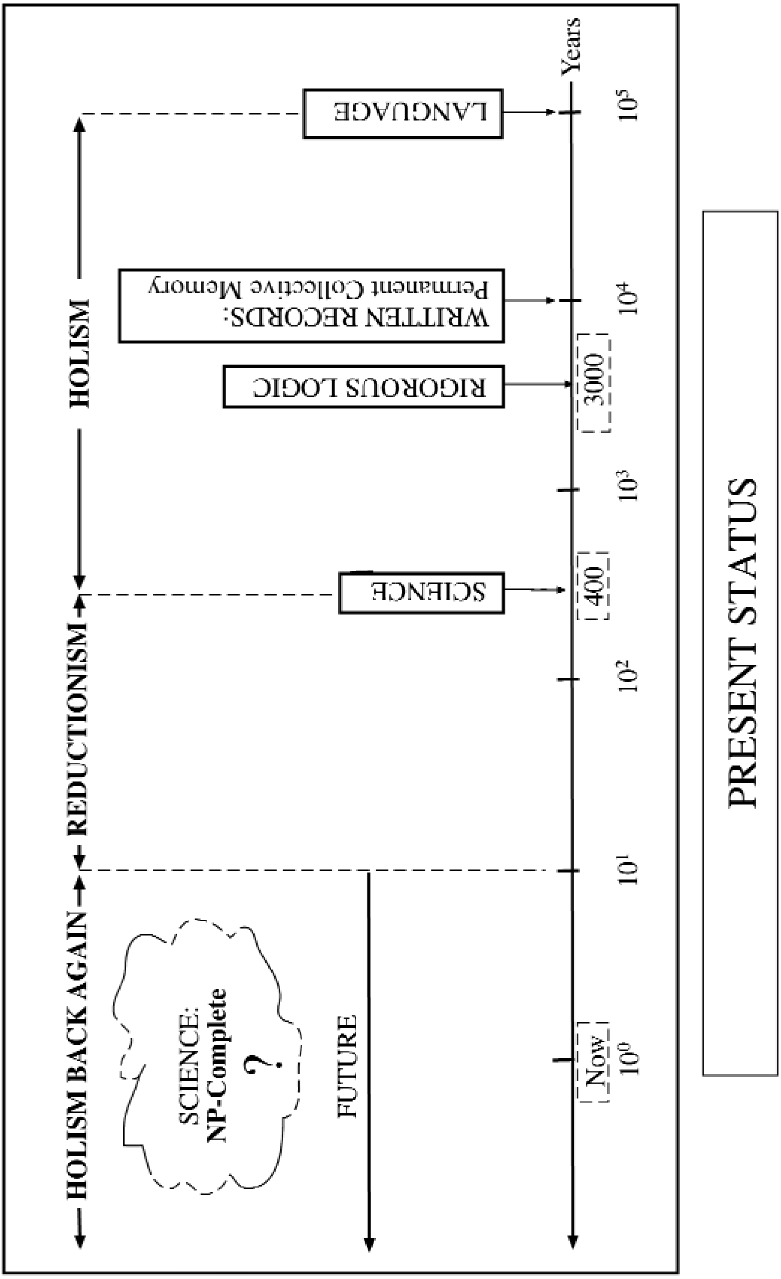

PRESENT STATUS
Years
10^5
LANGUAGE
WRITTEN RECORDS:
Permanent Collective Memory
10^4
RIGOROUS LOGIC
3000
10^3
SCIENCE
400
10^2
10^1
10^0
Now
HOLISM
REDUCTIONISM
HOLISM BACK AGAIN
SCIENCE:
NP-Complete
?
FUTURE

Figure 12.

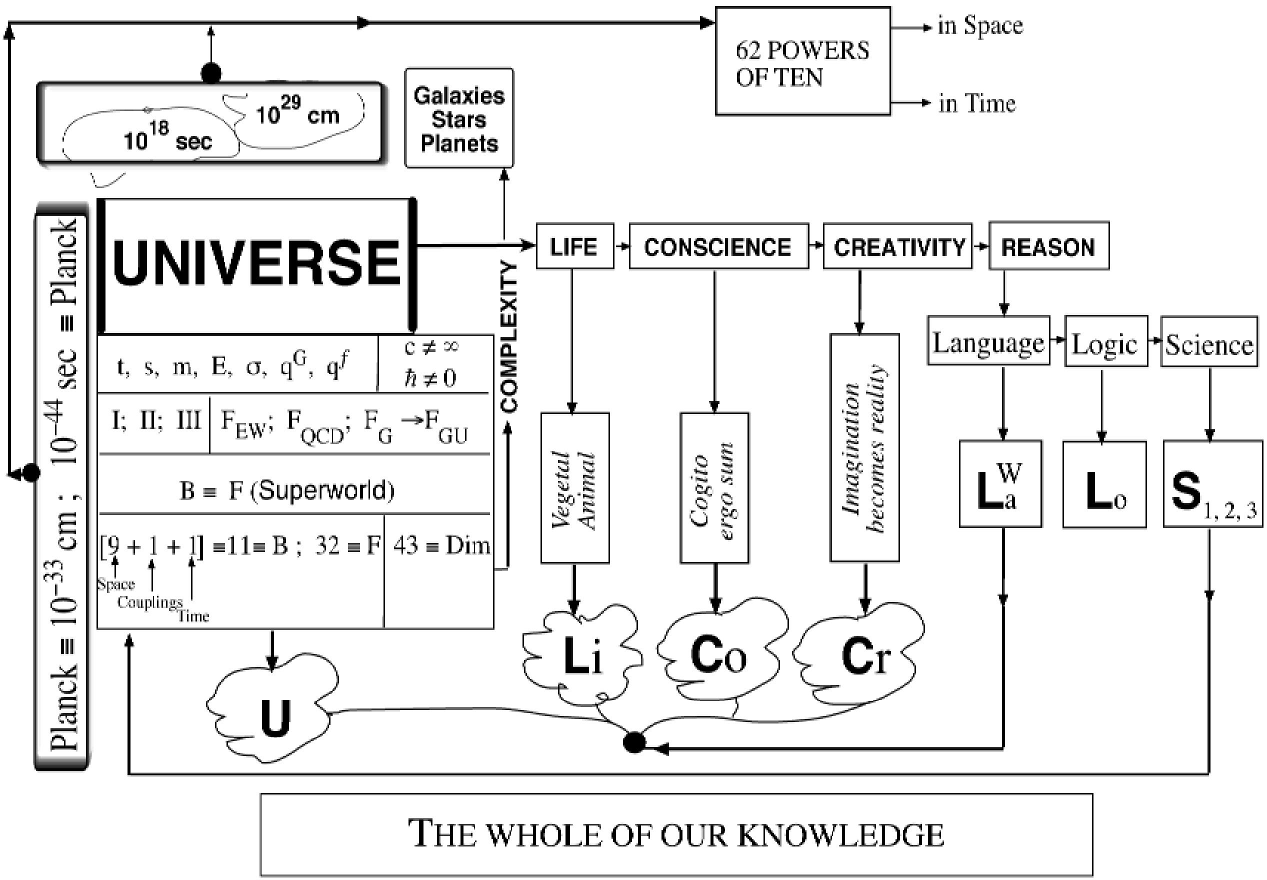

62 POWERS OF TEN
in Space
in Time
10^{29} cm
10^{18} sec
Galaxies Stars Planets
UNIVERSE
Planck ≡ 10^{-33} cm ; 10^{-44} sec ≡ Planck
t, s, m, E, σ, q^G, q^f
$c \neq \infty$
$\hbar \neq 0$
I; II; III
F_{EW}; F_{QCD}; $F_G \to F_{GU}$
B ≡ F (Superworld)
[9 + 1 + 1] ≡11≡ B ; 32 ≡ F
43 ≡ Dim
Space
Couplings
Time
COMPLEXITY
LIFE
CONSCIENCE
CREATIVITY
REASON
Vegetal Animal
Cogito ergo sum
Imagination becomes reality
Language
Logic
Science
L_a^W
L_o
$S_{1,2,3}$
U
Li
Co
Cr
THE WHOLE OF OUR KNOWLEDGE

Figure 13.

Let me point out that Science is the consequence of us being the only form of living matter endowed with Reason, from where the sequence of Language–Logic–Science has originated [10]. The time-sequence of Language–Logic–Science is shown in figure 14.

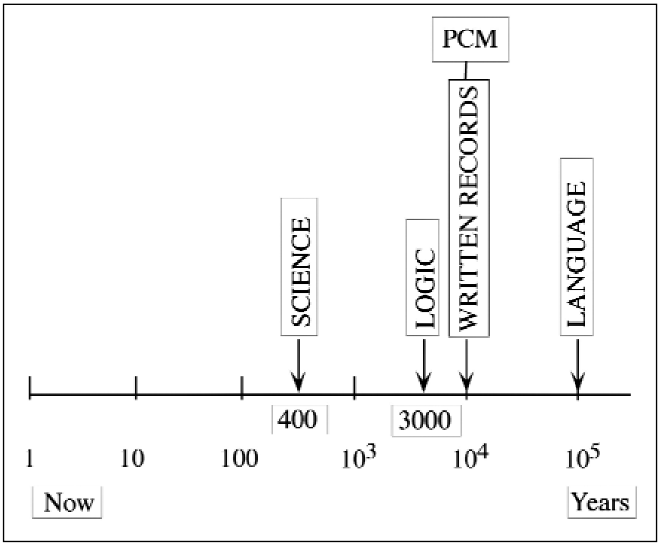


Figure 14. The Time-Sequence of Language – Logic – Science.

How can we interpret the fact that the greatest achievements of Science have always been originated by totally unexpected events? *Why? Answer*: Because the fellow who created the world is smarter than all of us, including scientists, mathematicians, philosophers.

9. References

[1] *Complexity exists at the Fundamental Level*
Proceedings of the 2004–Erice Subnuclear Physics School to be published by World Scientific (2006).

[2] *The Landscape and its Physics Foundations*
How String Theory Generates the Landscape
L. Susskind
Proceedings of the 2006–Erice Subnuclear Physics School to be published by World Scientific.

[3] *Complexity and Landscape in String Theory*
M.R. Douglas and F. Denef
Proceedings of the 2006–Erice Subnuclear Physics School to be published by World Scientific.

[4] *Complexity at the Fundamental Level*
A. Zichichi
presented at:

– *International Conference on 'Quantum [un]speakables' in Commemoration of John S. Bell, International Erwin Schrödinger Institut (ESI), Universität Wien (Austria), November 2000,* 'John Bell and the Ten Challenges of Subnuclear Physics'.

– 40th Course of the International School of Subnuclear Physics, Erice (Italy), September 2002, '*Language Logic and Science*'.

– 31st, 32nd and 33th Course of the International School of Solid State Physics, Erice (Italy), July 2004, '*Complexity at the Elementary Level*'.

– 42nd International School of Subnuclear Physics, Erice (Italy), August-September 2004, '*Complexity at the Elementary Level*'.

– Trinity College, Dublin (Ireland), February 2005, '*Complexity at the Elementary Level*'.

– Department of Physics, University of Padova (Italy), March 2005, '*Complexity at the Elementary Level*'.

– 43th Course of the International School of Subnuclear Physics, Erice (Italy), September 2005, '*Complexity at the Elementary Level*'.

– Italian Physics Society (SIF) XCI Annual National Congress, University of Catania (Italy), September 2005, '*Complexity at the Elementary Level*'.

– Desy, Hamburg, November 2005, '*Complexity at the Fundamental Level*'.

– 44th Course of the International School of Subnuclear Physics, Erice (Italy), September 2006, '*Complexity at the Fundamental Level*'.

[5] *The Logic of Nature and Complexity*
A. Zichichi
presented at:

– Pontificia Academia Scientiarum, The Vatican, Rome (Italy), November 2002, *'Scientific Culture and the Ten Statements of John Paul II'; 'Elements of Rigour in the Theory of Evolution'*.

– The joint Session of:
6th Course of the International School of Biological Magnetic Resonance; Erice (Italy), July 2003, *'Language Logic and Science'*.

– 2nd Workshop on Science and Religion of the Advanced School of History of Physics; Erice (Italy), July 2003, *'Language Logic and Science'*.

– 10th Workshop of the International School of Liquid Crystals; Erice (Italy), July 2003, *'Language Logic and Science'*.

– International School on Complexity, 1st Workshop on Minimal Life, Erice (Italy), December 2004, *'Evolution and Complexity at the Elementary Level'*.

[6] *Complexity and New Physics*
A. Zichichi
presented at:

– INFN-Alice Meeting, University of Catania (Italy), January 2005, *'Complexity at the Elementary Level'*.

– INFN Eloisatron Project 'The 1st Physics ALICE Week', Erice (Italy), December 2005, *'Complexity and New Physics with ALICE'*.

– 50th Anniversary of INFN Bologna - ALICE Week, Bologna (Italy), June 2006, *'Complexity at the Fundamental Level'*.

[7] *Complexity and Planetary Emergencies*
A. Zichichi
presented at:

– 27th Sessions of the International Seminars on Planetary Emergencies, Erice (Italy), August 2002, *'Language, Logic and Science'*.

– 28th Sessions of the International Seminars on Planetary Emergencies, Erice (Italy), August 2003, *'Language Logic and Science, Evolution and Planetary Emergencies'*.

– 36th Sessions of the International Seminars on Planetary Emergencies, Erice (Italy), August 2006, *'Complexity and Planetary Emergencies'*.

[8] *Virtual History: Alternatives and Counterfactuals*
Niall Ferguson ed., Basic Books, New York (1999).

[9] *The Logic of Nature, Complexity and New Physics: From Quark-Gluon Plasma to Superstrings, Quantum Gravity and Beyond*
Proceedings of the 2006–Erice Subnuclear Physics School to be published by World Scientific.

[10] *Language, Logic and Science*
A. Zichichi, proceedings of the 26th Session of the International Seminar on Nuclear War and Planetary Emergencies, Erice 18-26 August 2002 (World Scientific, 2003).

The References [4-7] refer to the various occasions where I have presented papers on highly specialized topics and discussed the connection of these topics with Complexity. The title on the upper part refers to the connection with Complexity while the specialized topic is reported in the detailed references.

Session II

GEOSCIENCES AND ENVIRONMENTAL EVOLUTION

CHANGE OF SCALING BEFORE EXTREME EVENTS IN COMPLEX SYSTEMS

VLADIMIR KEILIS-BOROK,[1,2] ALEXANDRE SOLOVIEV,[1,3]
ANDREI GABRIELOV,[4] ILIA ZALIAPIN[1,5]

1. Introduction

Critical transitions. Natural and human-made complex systems persistently generate critical transitions – rare extreme events, also known as disasters, crises etc. Predictive understanding of critical transitions is commonly regarded as one of the major unsolved problems of basic science.

The dramatic practical aspect of that problem is disaster preparedness. Global population, economy, and environment become increasingly vulnerable to a throng of disasters, already threatening the very survival and sustainability of our civilization (Science for Survival and Sustainable Development, 2000; G8-UNESCO World Forum on 'Education, Innovation and Research: New Partnership for Sustainable Development', 2007).

This study explores a specific observable phenomenon, preceding critical transitions. It is a change of scaling – the size distribution of events comprising a process considered. Scaling itself is a staple of studying complexity. However premonitory change of scaling has been found so far only in seismicity and in other forms of multiple fracturing, with major failures (e.g.

[1] International Institute of Earthquake Prediction Theory and Mathematical Geophysics, Russian Academy of Sciences, Moscow, Russia.

[2] Institute of Geophysics and Planetary Physics and Department of Earth and Space Sciences, University of California, Los Angeles, USA.

[3] The Abdus Salam International Centre for Theoretical Physics, Trieste, Italy

[4] Departments of Mathematics and Earth and Atmospheric Sciences, Purdue University, West Lafayette, USA.

[5] University of Nevada, Reno, USA.

strong earthquakes) for critical transitions (e.g., Smith, 1986; Main *et al.*, 1989, 1992; Henderson *et al.*, 1992, 1994; Narkunskaya and Shnirman, 1994; Rotwain *et al.*, 1997; Wiemer and Wyss, 1997; Wyss and Wiemer, 2000; Burroughs and Tebbens, 2002; Amitrano, 2003; Zaliapin *et al.*, 2003).

Here we explore universality of this phenomenon, by uniform analysis of a variety of data: observations relevant to economic recessions, surges of unemployment, and homicides surges; and results of mathematical modeling. Altogether this seems to open the promising line of research in predicting critical transitions.

Next we outline the background of this study.

Predictability. Complex systems are not predictable in the Laplacean sense, with accuracy limited only by accuracy of data and theory. However, after a coarse graining, in a not-too-detailed scale, such systems exhibit regular behaviour patterns and become predictable, up to the limits. The holistic approach, 'from the whole to details' opens a possibility to overcome the complexity itself and the chronic imperfection of data as well (Farmer and Sidorowich, 1987; Ma *et al.*, 1990; Kravtsov, 1993; Gell-Mann, 1994; Holland, 1995; Kadanoff, 1976; Keilis-Borok and Soloviev, 2003).

Premonitory patterns. Dynamics of a complex system include a variety of observable processes; they show premonitory patterns emerging as an extreme event approaches (e.g., Keilis-Borok and Malinovskaya, 1964; Gabrielov *et al.*, 1986; Keilis-Borok, 1990; Newman *et al.*, 1995; Sammis *et al.*, 1996; Keilis-Borok and Shebalin, 1999; Sornette, 2000; Keilis-Borok and Soloviev, 2003). Note that while the targets of prediction – rare extreme events - have *low probability to occur but large impact on a system*, premonitory patterns are formed by more frequent events with *high probability to occur but low impact.*

Numerical modeling and data analysis demonstrated the *dual nature* of premonitory patterns: they are partly universal, as befits complexity, and partly process-specific. Change of scaling, explored here, is one of such patterns.

Difficulty and challenge of the predictability problem is also common for the frontiers of basic science: this is the absence of a complete set of fundamental equations governing formation of extreme events. A.N. Kolmogorov wrote:

> It became clear for me that it is unrealistic to have a hope for the creation of a pure theory [of the turbulent flows of fluids and gases] closed in itself. Due to the absence of such a theory we have to rely upon the hypotheses obtained by processing of the experimental data...

Pure theory is even less complete for dynamics of solid Earth (Keilis-Borok and Soloviev, 2003; Press, 1965) and – particularly – for socio-economic and political systems. Hence the bane of data processing – large number of adjustable parameters free for data-fitting.

New analytical framework. A possibility to reduce that number is recently provided by an exactly solvable model of diffusion with branching (Gabrielov *et al.*, 2007). This model analytically defines premonitory change of scaling and other premonitory patterns – *clustering in space-time* and *long-range correlations*. They have been found in spatially extended systems, real and numerically simulated (e.g., Keilis-Borok, 1990; Romanowicz, 1993; Press and Allen, 1995; Sornette and Sammis, 1995; Aki, 1996; Bowman *et al.*, 1998; Pollitz *et al.*, 1998; Keilis-Borok and Shebalin, 1999; Turcotte *et al.*, 2000). The branching diffusion model defines these patterns through a small number of common control parameters, thus liberating predictability studies from a major millstone.

2. Data Analysis

2.1. *Definitions*

– Prediction is targeted at extreme point events with occurrence times T_e, $e = 1, 2, \ldots$

– Premonitory patterns are looked for in an observable time process *S(t)*, hypothetically containing such patterns. In many problems this process is defined as a time series (t_i, m_i, g_i), $i = 1, 2, \ldots$ Here t is the time of the event, $t_i, \leq t_{i+1}$; m is its size (often given in logarithmic scale), g stands for additional parameters that might be indicated (e.g., vector of coordinates of earthquake's hypocenter).

– Scaling of a process *S(t)* is a function $N^{S}(m)$ – the number of events of the size $\geq m$. We consider its normalized form equivalent to probabilistic distribution function: $P^{S}(m) = N^{S}(m)/\check{N}^{S}$. Here $\check{N}^{S}$ is the total number of events considered; by definition this is the ordinate of the left end of that curve (at minimal m).

– In many problems the data consist of some average characteristics of system's behavior. For the socio-economic crises considered here these characteristics are monthly indicators. In such a case scaling is determined for the change of the indicator's trend; its definition follows. *f(t)* – a monthly indicator. $W^f(t/q)=K^f(t/q)(t-q)\ +\ B^f(t/q),\ 0<t<q$. This is the local linear least-squares regression of *f(t)* within the sliding time window (*t-q, t*). $S^f(t/s,u)\ =\ K^f(t/s)\ –\ K^f(t-s/u)$ – an 'event': the change of the trend $K^f(t/q)$ between consecutive intervals: current $(t-s,\ t)$ and previous $(t-s-u,\ t-s)$ intervals. Time and, accordingly, parameters *s*, *u* are the integers, measured in the number of months. Size *m* of events, for which the scaling is determined, is the absolute value of S^f.

– Time considered is divided into periods of three kinds, as shown in Fig. 1 (see page 240). To explore premonitory change of scaling we compare functions *P*(*M*), and number of events $\check{N}$ in the periods N and D. In the subsequent text lower indexes identify these periods (e.g., $\check{N}_{\mathrm{N}}$, $\check{N}_{\mathrm{D}}$).

2.2. *Point of Departure: Strong Earthquakes*

Three examples of seismicity analysis are shown in Fig. 2 (see page 240). *Critical transitions* ('prediction targets') are the main shocks with magnitude $M\geq6.4$; here *M* is the logarithmic measure of energy released by an earthquake. *Size distribution P(m)* is probability that the size of an event is $\geq m$; total number of events is $\check{N}_{\mathrm{N}}=277$and $\check{N}_D=255$ for the periods N and D respectively; *a*: events are individual main shocks; measure of size *m* is their magnitude; *b, c*: events are clusters of aftershocks formed around individual main shocks (Keilis-Borok *et al.*, 1980; Molchan *et al.*, 1990); measure of cluster's size is number of aftershocks not weighted (*b*) or weighted (*c*) by their magnitudes.

2.3. *Socio-Economic Crises*

Three examples of data analysis are shown in Fig. 3 (see page 240). *Critical transitions* ('prediction targets') are the starting points of a respective crisis. *Size distribution P(m)* is probability that the size of an event is $\geq m$. Event is the change of a monthly indicator considered: industrial production before recessions (*a*) and unemployment surges (*b*); and monthly rates of lesser crimes – assaults with firearms – before homicide surge (*c*). Total number of events in periods N and D: $\check{N}_{\mathrm{N}}=\check{N}_{\mathrm{D}}=62$ (*a*); $\check{N}_{\mathrm{N}}=24$, $\check{N}_{\mathrm{D}}=44$ (*b*); and $\check{N}_{\mathrm{N}}=21$, $\check{N}_{\mathrm{D}}=28$ (*c*).

2.4. *Modeling*

Similar premonitory change of scaling has been found in a variety of models

– Models of inverse, direct, and colliding cascades (e.g., Allègre *et al.*, 1982, 1995; Narkunskaya and Shnirman, 1994; Gabrielov *et al.*, 2000; Zaliapin *et al.*, 2003).

– Laboratory experiments with fracturing of rocks and metals (Rotwain *et al.*, 1997).

– Models of tectonic blocks and faults (Soloviev and Ismail-Zadeh, 2003).

– Branching diffusion model (Gabrielov *et al.*, 2007).

3. Discussion

– We followed here the 'premonitory patterns' approach in which prediction is targeted at the rare extreme point events. This approach is complementary to classical Kolmogoroff-Wiener prediction, targeted at extrapolation of a whole process, i.e. mainly at the medium or small events.

– That approach is complementary also to cause-and-effect analysis. Extreme events and premonitory patterns often are the parallel manifestations of evolution of the complex system.

– Furthermore premonitory patterns might predict not an extreme event *per se*, but the system's destabilization making it ripe for an extreme event; its triggering then becomes close to inevitable, not requiring a particularly strong impact.

– Transition to predicting individual extreme events comprises at least the following further problems:

Parametrization of premonitory change of scaling, sensitivity analysis, and optimization (Molchan, 2003).

Similar analysis for clustering (e.g., Keilis-Borok *et al.*, 1980) and correlation range (e.g., Shebalin, 2006).

Acknowledgements

This study was supported by the EC Project 'Extreme events: Causes and consequences (E2-C2), Contract No 12975 (NEST)'.

REFERENCES

Aki, K. Scale dependence in earthquake phenomena and its relevance to earthquake prediction. *Proc. Natl. Acad. Sci.* USA, 93, 3740–3747, 1996.

Allègre, C.J., Le Mouël, J.-L., and Provost, A. Scaling rules in rock fracture and possible implications for earthquake prediction. *Nature*, 297, 47-49, 1982.

Allègre, C.J., Le Mouël, J.-L., Chau, H.D. and Narteau, C. Scaling organization of fracture tectonics (SOFT) and earthquake mechanism. *Phys. Earth Planet. Inter.*, 92, 215-233, 1995.

Amitrano, D. Brittle-ductile transition and associated seismicity: Experimental and numerical studies and relationship with the *b* value. *J. Geophys. Res.*, 108, 2044, doi:10.1029/2001JB000680, 2003.

Bowman, D.D., Ouillon, G., Sammis, G.G., Sornette, A., and Sornette, D. An observational test of the critical earthquake concept. J. Geophys. Res., 103, 24359-24372, 1998.

Burroughs, S.M. and Tebbens, S.F. The upper-truncated power law applied to earthquake cumulative frequency-magnitude distributions: evidence for a time-independent scaling parameter. *Bull. Seismol. Soc. Am.*, 92, 2983-2993, 2002.

Farmer, J.D. and Sidorowich, J. Predicting chaotic time series. *Phys. Rev. Lett*. 59, 845, 1987.

G8-UNESCO World Forum on “Education, Innovation and Research: New Partnership for Sustainable Development”, 2007, 10-12 May, Trieste, Italy, http://g8forum.ictp.it/

Gabrielov, A., Dmitrieva, O.E., Keilis-Borok, V.I,, Kossobokov, V.G., Kuznetsov, I.V., Levshina, T.A., Mirzoev, K.M., Molchan, G.M., Negmatullaev, S.Kh., Pisarenko, V.F., Prozoroff, A.G., Rinehart, W., Rotwain, I.M., Shebalin, P.N., Shnirman, M.G., and Shreider, S.Yu. Algorithm of Long-term Earthquakes’ Prediction. Centro Regional de Sismología para América del Sur, Lima (Peru), 1986.

Gabrielov, A.M., Zaliapin, I.V., Newman, W.I., and Keilis-Borok, V.I. Colliding cascade model for earthquake prediction. *Geophys. J. Int.*, 143(2), 427-437, 2000.

Gabrielov, A., Keilis-Borok, V., and Zaliapin, I. Predictability of extreme events in a branching diffusion model. arXiv:0708.1542 [nlin.AO], 2007.

Gell-Mann, M. The Quark and the Jaguar: Adventures in the Simple and the Complex. Freeman and Company, New York, 1994.

Henderson, J., Main, I., Meredith, P., and Sammonds, P. The evolution of seismicity at Parkfield: observation, experiment and a fracture-mechanical interpretation. *J. Struct. Geol.*, 14, 905-913, 1992.

Henderson, J., Main, I.G., Pearce, R.G., and Takeya, M. Seismicity in north-eastern Brazil: fractal clustering and the evolution of the *b* value. *Geophys. J. Int.*, 116, 217-226, 1994.

Holland, J.H. Hidden Order: How Adaptation Builds Complexity. Addison-Wesley, Reading (Mass), 1995.

Kadanoff, L.P. Scaling, universality and operator algebras. In: Domb, C. and Green, M.S. (eds.) Phase Transitions and Critical Phenomena, Vol. 5a, 1976.

Keilis-Borok, V.I. and Malinovskaya, L.N. One regularity in the occurrence of strong earthquakes. J. Geophys. Res., 69, 3019–3024, 1964.

Keilis-Borok, V.I., Knopoff, L. and Rotwain I.M. Bursts of aftershocks, long-term precursors of strong earthquakes. Nature, 283, 258–263, 1980.

Keilis-Borok, V.I. (ed.) Intermediate-Term Earthquake Prediction: Models, Algorithms, Worldwide Tests. *Phys. Earth Planet. Inter.*, 61(1-2), special issue, 1990.

Keilis-Borok, V.I. and Shebalin, P.N. (eds.) Dynamics of Lithosphere and Earthquake Prediction. *Phys. Earth Planet. Inter.*, 111(3-4), special issue, 1999.

Keilis-Borok, V.I. and Soloviev, A.A. (eds.) *Nonlinear Dynamics of the Lithosphere and Earthquake Prediction*, Springer-Verlag, Berlin-Heidelberg, 2003.

Kravtsov, Yu.A. (ed.) *Limits of Predictability*. Springer-Verlag, Berlin-Heidelberg, 1993.

Ma, Z., Fu, Z., Zhang, Y., Wang, C., Zhang, G., and Liu, D. Earthquake Prediction: Nine Major Earthquakes in China. Springer-Verlag, New York, 1990.

Main, I.G., Meredith, P.G., and Jones, C. A reinterpretation of the precursory seismic *b*-value anomaly from fracture mechanics. *Geophys. J. Int.*, 96, 131-138, 1989.

Main, I.G., Meredith, P.G., and Sammonds, P.R. Temporal variations in seismic event rate and *b*-values from stress corrosion constitutive laws. *Tectonophysics*, 211, 233-246, 1992.

Molchan, G.M., Dmitrieva, O.E., Rotwain, I.M. and Dewey, J. Statistical analysis of the results of earthquake prediction, based on burst of aftershocks. *Phys. Earth Planet. Inter.*, 61, 128–139, 1990.

Molchan, G.M. Earthquake Prediction Strategies: A Theoretical Analysis. In: Keilis-Borok, V.I. and Soloviev, A.A. (eds.) *Nonlinear Dynamics of the Lithosphere and Earthquake Prediction*. Springer-Verlag, Berlin-Heidelberg, pp. 209-237, 2003.

Narkunskaya, G.S. and Shnirman, M.G. On an algorithm of earthquake prediction. In: Chowdhury, D.K. (ed.) *Computational Seismology and Geodynamics*, Vol. 1, pp. 20-24, AGU, Washington, D.C., 1994.

Newman, W.I., Turcotte, D.L., and Gabrielov, A.M. Log-periodic behavior of a hierarchical failure model with application to precursory seismic activation. *Phys. Rev. E*, 52, 4827-4835, 1995.

Pollitz, F.F., Burgmann, R., and Romanowicz, B. Viscosity of oceanic asthenosphere inferred from remote triggering of earthquakes. *Science*, 280, 1245-1249, 1998.

Press, F. (ed.) Earthquake Prediction: A Proposal for a Ten Year Program of Research. Ad Hoc Panel on Earthquake Prediction, White House Office of Science and Technology, Washington, D.C., 134 pp., 1965.

Press, F. and Allen, C. Patterns of seismic release in the southern California region. *J. Geophys. Res.*, 100(B4), 6421–6430, 1995.

Romanowicz, B. Spatiotemporal patterns in the energy-release of great earthquakes. *Science*, 260, 1923-1926, 1993.

Rotwain, I., Keilis-Borok, V. and Botvina, L. Premonitory transformation of steel fracturing and seismicity. *Phys. Earth Planet. Inter.*, 101, 61-71, 1997.

Sammis, C.G., Sornett, D., and Saleur, H. Complexity and earthquake forecasting. In: Rundle, J.B., Turcotte, D.L., and Klein, W. (eds.) SFI Studies in the Science of Complexity, Vol. XXV, Addison-Welsey, Reading (Mass), 1996.

Science for Survival and Sustainable Development. The Proceedings of the Study-Week of the Pontifical Academy of Sciences, 12-16 March 1999. Pontificiae Academiae Scientiarvm Scripta Varia, 98, Vatican City, 2000.

Shebalin, P. Increased correlation range of seismicity before large events manifested by earthquake chains. Tectonophysics, 424, 335-349, 2006.

Smith, W.D. Evidence for precursory changes in the frequency-magnitude *b*-value, *Geophys. J. Int.*, 86, 815-838, 1986.

Soloviev, A. and Ismail-Zadeh, A. Models of dynamics of block-and-fault systems. In: Keilis-Borok, V.I. and Soloviev, A.A. (eds.) Nonlinear Dynamics of the Lithosphere and Earthquake Prediction. Springer-Verlag, Berlin-Heidelberg, pp. 71-139, 2003.

Sornette, D., Sammis, C.G. Complex critical exponents from renormalization group theory of earthquakes: Implications for earthquake predictions. *J. Phys. I France*, 5, 607–619, 1995.

Sornette, D. *Critical Phenomena in Natural Sciences: Chaos, Fractals, Self-organization, and Disorder. Concept and Tools*. Springer, Berlin, 2000.

Turcotte, D.L., Newman, W.I., and Gabrielov, A. A statistical physics approach to earthquakes. In: Geocomplexity and the Physics of Earthquakes. *Am. Geophys Un.*, Washington, DC, 2000.

Wiemer, S. and Wyss, M. Mapping the frequency-magnitude distribution in asperities: An improved technique to calculate recurrence times? *J. Geophys. Res.*, 102, 15115-15128, 1997.

Wyss, M. and Wiemer, S. Change in the probability for earthquakes in Southern California due to the Landers magnitude 7.3 earthquake. *Science*, 290, 1334-1338, 2000.

Zaliapin, I., Keilis-Borok, V. and Ghil, M. A Boolean delay model of colliding cascades. II: Prediction of critical transitions. *J. Stat. Phys.*, 111, 839-861, 2003.

GLOBAL WARMING SCIENCE: PREDICTIONS, SURPRISES AND UNCERTAINTIES

VEERABHADRAN RAMANATHAN

Preface

Much too much has been written about climate change. Yet, very few of these articles elaborate on how the field has rapidly evolved over the last few decades; how discoveries are made to understand and keep up with the rapid warming of the planet; and above all, how climate science is conducted. These are the issues that I am going to dwell on in this lecture, since the title of this meeting at the Pontifical Academy of Sciences encourages me to do so.

The first scholarly and quantitative work on the greenhouse effect of carbon dioxide was done nearly one hundred years ago by Svante Arrhenius, the Swedish Nobel chemist. Arrhenius (1896) developed a simple mathematical model for the transfer of radiant energy through the atmosphere-surface system and solved it analytically to show that a doubling of the atmospheric concentration would lead to a warming of the surface by as much as 4 to 5K. Since then there has been a tremendous amount of work on the science of global warming, culminating in the now famous, Intergovernmental Panel on Climate Change (IPCC) reports. The next major development was the idea of Lorenz proposed in the 1970s, in which he used simple but elegant mathematical models of the dynamics of the atmosphere to reveal the fundamental unpredictability of weather and thus laid a solid foundation for the science of Chaos. Arrhenius' greenhouse effect model of climate change and Lorenz's Chaos model of weather and climate prediction provides two strong book-ends to the book on climate change, parts of which have been written during the last century.

In this lecture I would like to focus on the scientific underpinnings of the link between greenhouse gases and global warming. I will describe a personal journey in trying to understand how human activities are modifying the earth's climate and environment. The journey took some major twists and turns through adventurous paths and I have made an attempt to describe how one question led to another and connect the various findings over a period of 35 years. Over time I also had to acquire and develop new tools including climate models, satellite observations, field experiments and lightweight unmanned aircraft. Ultimately these made me realize the seriousness of the climate change situation, threatening the water and food security of the planet. This realization has inadvertently taken me down the path of proposing practical solutions for mitigating unprecedented climate changes.

Since this is a lecture, I have not attempted to refer to many valuable references by groups other than mine. I refer the audience to many insightful surveys, including the historical background by Le Truet and Somerville (2007) in the IPCC-AR4 (2007) report, as well as to the entire IPCC report.

I. Predictions of the Anthropogenic Forcing of Climate

Inadvertent Modification of the Atmosphere

When we look up to the sky, the atmosphere seems enormous and limitless. But it is really a thin shell of gases, particles and clouds surrounding the planet (Figure 1, see p. 241). It is in this thin shell that we are dumping several billion tons of pollutants each year. The major sources of this pollution include fossil fuel combustion for power generation and transportation (Figs 2a and 2b, see p. 241); cooking with solid fuels (Fig 2c, see p. 241); and burning of forests and savannah (Figure 2d, see p. 241). The ultimate by-product of all forms of burning is the emission of the colourless gas, carbon dioxide (CO_2). Now let us consider the following question: *Why should a person in my ancestral village in S India (Fig 2, bottom left panel) worry about the pollutant emitted by someone travelling in a car in the US (top right hand panel)? Likewise, why should we in the US worry about the cooking habits of people in my ancestral village?*

The answer is simple. Every part of the world is connected with every other part through fast atmospheric transport. For example Figure 3 (see p. 241) shows a one-day snapshot of the synoptic distribution of water vapour in the atmosphere as simulated by a climate model. This climate model is called the community climate model (CCM) developed by the National Center for Atmospheric Research in Boulder, Colorado. I was part of a team of 5 scientists who developed the first version of this model in the early 1980s (Pitcher *et al.*, 2003; Ramanathan *et al.*, 2003). It was then referred to as CCM0. The simulation shown in Fig. 3 is from CCM3, the fourth version of the model published recently (Kiehl *et al.*, 2004). The result shown for water vapour is basically a solution for a highly non-linear system of dynamical-thermodynamical equations. Most three-dimensional climate models solve for the climate by simulating individual weather systems like the ones shown in Fig. 3, every one hour for hundreds of years. The synoptic pattern clearly shows how air parcels can extend thousands of kilometres across from East Asia into N America; from N America across the Atlantic into Europe; from S Asia into E Asia; from Australia into the Antarctic and so on. Aircraft and satellite data clearly reveal that, within a week, emissions, be it from Asia, N America or Africa are transported half way around the world into trans-oceanic and trans-continental plumes.

The lifetime of a CO_2 molecule is of the order of decades to century. This is more than sufficient time for the billions of tons of manmade CO_2 to uniformly cover the planet like a blanket. Do we have evidence for this colourless blanket? Figure 4 (see p. 242) shows a time series of CO_2 in the air collected by Keeling of the Scripps Institution of Oceanography, from the now famous Mauna Loa observatory. The steady increase is basically the icon for most discussions about climate change. Clearly CO_2 concentration in the air is increasing and this increase is due to human activities. This CO_2 increase is observed no matter where we make these observations, attesting to the fact that, basically, carbon dioxide has surrounded the planet like a blanket, all the way from the surface up to 100 km.

The question is, why should we worry about this colourless gaseous blanket?

The Climate System: Basic Drivers

First, I have to provide a brief background of how the climate system works.

Fundamentally, the incident solar radiation drives the climate system as well as life (Figure 5, see p. 242). About 30% of the incoming solar energy is reflected back to space. The balance of 70% is absorbed by the surface-atmosphere system. This energy heats the planet and the atmosphere. As the surface and the atmosphere warms, it gives off the energy as infrared radiation, also referred to as 'heat radiation'. So this process of the net incoming (downward solar energy-reflected) solar energy warming the system and the outgoing heat radiation from the warmer planet escaping to space goes on until the two components of the energy are in balance. In an average sense, it is this radiation energy balance that provides a powerful constraint for the global average temperature of the planet.

The Greenhouse Effect: The CO_2 Blanket

On a cold winter night, a blanket keeps the body warm not because the blanket gives off any energy. Rather, the blanket traps the body heat, preventing it from escaping to the colder surroundings. Similarly, the CO_2 blanket traps the heat radiation given off by the planet. The trapping of the heat radiation is dictated by quantum mechanics. The two oxygen atoms in CO_2 vibrate with the carbon atom in the center and the frequency of this vibration coincides with some of the infrared wavelengths of the heat radiation.

When the frequency of the heat radiation from the earth's surface and the atmosphere coincides with the frequency of CO_2 vibration, the radiation is absorbed by CO_2 and is converted to heat and is given back to the surface. As a result of this trapping, the outgoing heat radiation is reduced by increasing CO_2. Not as much heat is escaping to balance the net incoming solar radiation. There is excess heat energy in the planet, i.e., the system is out of energy balance. As CO_2 is increasing with time (Figure 4), the infrared blanket is becoming thicker, and the planet is accumulating this excess energy.

Global Warming: Getting Rid of the Excess Energy

How does the planet get rid of the excess energy? We know from the basic infrared laws of physics, the so-called Planck's black body radiation law, that warmer bodies emit more heat radiation. So, the system will get rid of this excess energy by warming and emitting more infrared radiation, until the excess energy trapped is given off to space and the surface-atmosphere system is in balance. That, in a nutshell, is the theory of the greenhouse effect and global warming. The rigorous mathematical modelling of this energy balance paradigm was originated by Arrhenius, but the proper accounting of the energy balance of the climate system containing the surface and the atmosphere had to await the work of Manabe and Wetherald in 1967.

CFCs: The Super-Greenhouse Gas

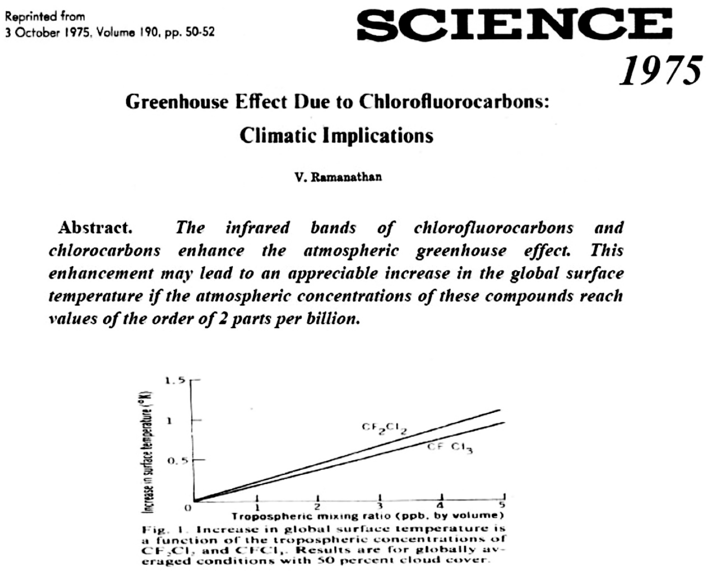
For nearly eighty years since Arrhenius' paper, climate scientists assumed that CO_2 was the main anthropogenic or manmade greenhouse gas (e.g., SMIC Report, 1971). For an entirely accidental reason, I stumbled onto the fact (Ramanathan, 1974) that there are other manmade gases, which on a per molecule basis are a thousand to more than ten thousand times stronger than the CO_2 greenhouse effect. Let me take the case of chlorofluorocarbons, or CFCs, used as refrigerants and propellants in deodorizers, drug delivery pumps, etc. These are purely synthetic gases. In 1974, Molina and Rowland published a famous paper in *Nature* (1974). In that paper, Molina and Rowland (1974) proposed that CFC11 and CFC12 (known then as Freon 11 and Freon 12), because of their century or longer lifetime, will build up in the atmosphere including the stratosphere. According to their theory, in the stratosphere, UV radiation from the sun will photo dissociate the CFCs and the chlorine atoms that are released will catalytically destroy ozone.

During 1974 to 1976, I was a National Research Council post doctoral fellow at the NASA Langley research center. The paper caught my attention. I was drawn to it, in part, because I had spent 2 years during 1965 to 1967 as an engineer in a refrigerator manufacturing company in Hyderabad, India. My job was to understand why CFCs leaked so quickly into the air from the sealed units, after delivery to the user! The other connecting event was my Ph D work during 1970 to 1973 in the US, where my thesis research dealt with the greenhouse effect of CO_2 in Mars and Venus. The CFC greenhouse effect finding was a question of connecting the dots.

One year after the Molina and Rowland paper, I published my findings in *Science* (Figure 6). What this work suggested, and it was met with disbelief, was that adding one molecule of CFC11 or CFC12 had the same effect as adding 10,000 molecules of CO_2! So suddenly human beings have synthesized and released this enormously powerful greenhouse gas. Why do CFCs have such a disproportionately large greenhouse effect? Before I can answer that, I need to explain the natural greenhouse effect.

Evidence for the Greenhouse Effect

Recall that, as I mentioned, adding a greenhouse gas would reduce the heat radiation escaping to space? How do we know this? Of course, in theory, basically all you need is quantum mechanics and radiation transfer equation to deduce the greenhouse effect rigorously, but I want to



Reprinted from
3 October 1975, Volume 190, pp. 50-52

SCIENCE

1975

Greenhouse Effect Due to Chlorofluorocarbons: Climatic Implications

V. Ramanathan

Abstract. *The infrared bands of chlorofluorocarbons and chlorocarbons enhance the atmospheric greenhouse effect. This enhancement may lead to an appreciable increase in the global surface temperature if the atmospheric concentrations of these compounds reach values of the order of 2 parts per billion.*

Increase in surface temperature (°K)
1.5
1
0.5
0
CF_2Cl_2
$CF\ Cl_3$
1
2
3
4
5
Tropospheric mixing ratio (ppb, by volume)

Fig. 1. Increase in global surface temperature is a function of the tropospheric concentrations of CF_2Cl_2 and $CFCl_3$. Results are for globally averaged conditions with 50 percent cloud cover.

Figure 6. CFC Warming.

demonstrate it through satellite observations. The satellite I used for this purpose is called the NIMBUS 4 which carried an infrared spectrometer looking down on earth with a scanning telescope; it was observing the emitted infrared radiation coming from the planet. I am showing (Figure 7) the emitted radiation data over the tropical ocean as a function of wavelength in the infrared region. I have superposed the black body radiation for two temperatures at 300K and 200K (the dashed lines). The 300K (27 °C) is close to the surface temperature of the tropical oceans. The 300K dashed curve is basically the IR radiation emitted by the ocean. Without an atmosphere, the radiation that would reach satellite altitudes would be exactly equal to the 300K black body curve. What NIMBUS 4 measured was substantially smaller; and the difference shown by the shaded region is the greenhouse effect; or metaphorically, the thickness of the blanket. Gases in the atmosphere reduce the outgoing heat radiation, because the atmosphere is colder than the ground. The emitted radiation (or alternately number of photons) increases strongly with temperature. Thus the warmer surface is emitting a lot more radiation. The colder atmosphere absorbs it and reemits it at the lower temperature, so the net effect is to reduce the radiation energy leaving the planet, just like the

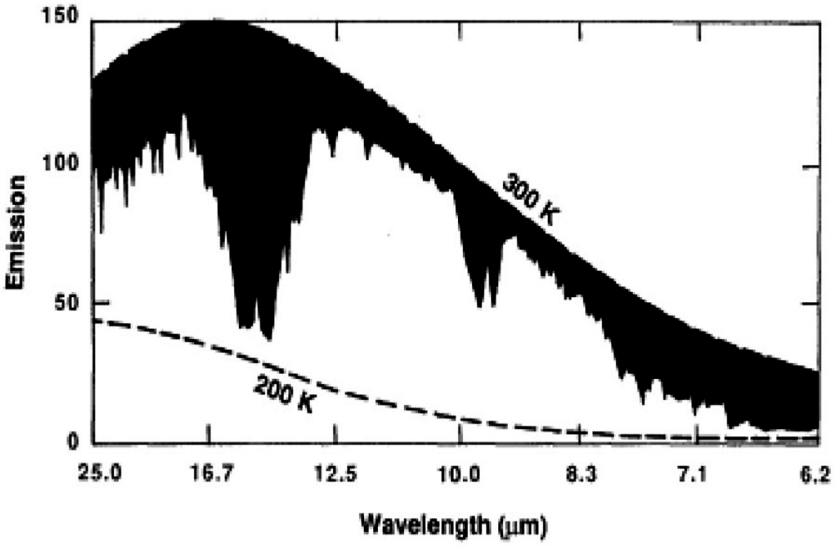


Figure 7. IRIS Spectra (adapted from Hanel *et al.*, 1972).

blanket inhibiting the flow of energy from your warm body to the colder environment.

We also see that the thickness is non-uniform and some (spectral) regions are much thicker than the others. This is because, there are many different greenhouse gases in the atmosphere and the greenhouse effect of each of these gases is different and in different spectral regions, as described next.

The Natural Greenhouse Effect: Nature's Blanket

The dip centered on the 15 μm (micrometer or micron) region is the greenhouse effect of CO_2; and that centered on 9.6 μm is due to ozone and most of the rest of the IR reduction is due to water vapor. Thus water vapour exerts the dominant greenhouse effect; CO_2 is next, followed by ozone and numerous other gases. The water vapour effect is completely natural. The CO_2 and ozone effects are also mostly natural. Later we will quantify the magnitude of the natural greenhouse effect and compare it with the manmade greenhouse effect. Figure 7 enables us to address the potency of the CFCs in reducing the outgoing IR.

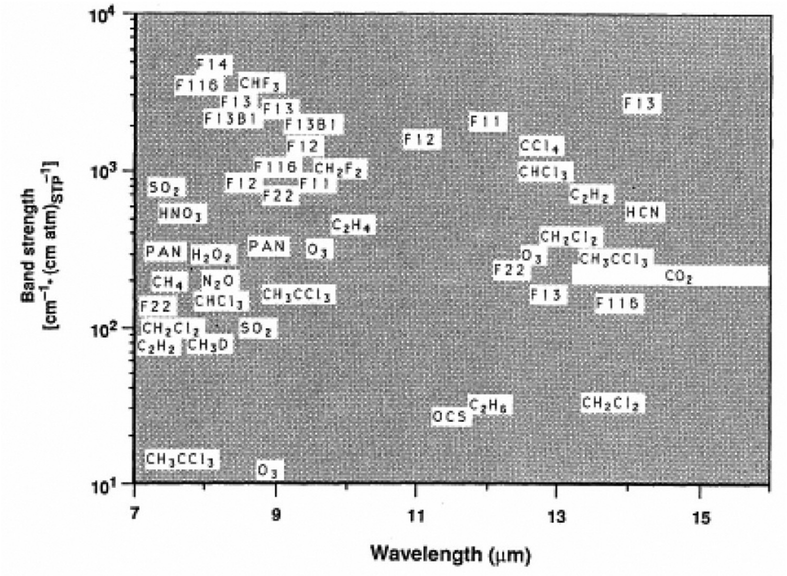


Figure 8. Ramanathan, 1988.

The Dirty Atmospheric Window

The blanket is thinnest in the 8 to 12 μm region, because the background atmosphere is mostly transparent in this spectral region and for this reason this region is called the atmospheric window. The background water vapour has very little absorption. What little you see is from the dimer, two water molecules dissociating to give this absorption. It is in this region that CFCs (and numerous other trace gases; Figure 8) absorb and emit radiation. In addition, the quantum mechanical efficiency (also knows as transition probability) of CFCs is about 3 to 6 times stronger than that due to CO_2. Lastly, the CFC concentrations are so low (part per billion or less) that their effect increases linearly with their concentration, whereas the CO_2 absorption is close to saturation since their concentration is about 300 000 times larger. So it's a lot harder for a CO_2 molecule to enhance the greenhouse effect than CFCs.

These three factors combine to make CFCs a super greenhouse gas. Numerous other manmade greenhouse gases (Figure 8) have similar strong absorption features in the window region, making the window a less transparent dirty window.

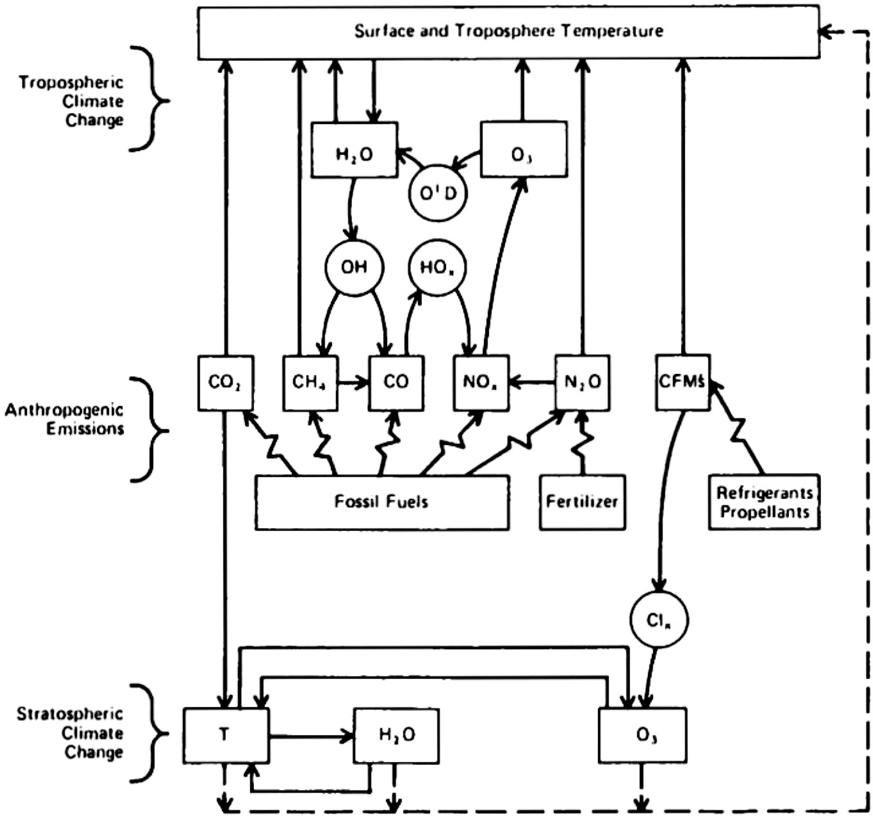


Figure 9. Ramanathan, 1980.

Climate-Chemistry Interaction

The independent discoveries of the CFC effect on the ozone chemistry and on the greenhouse effect coupled air-chemistry strongly with climate. Molina and Rowland's 1974 paper as well as Crutzen's (1972) earlier paper on the effect of nitrogen oxides (another pollutant) on the ozone layer, motivated me to look into the effect of stratospheric ozone on climate, because stratospheric ozone regulates the UV and visible solar radiation reaching the surface-troposphere (the first 10 to 16 km from the surface where the weather is generated) system; in addition, ozone, as discussed earlier, is a strong greenhouse gas. In 1976, I showed (Ramanathan *et al.*, 1976) that reducing ozone would cool not only the stratosphere (anticipated by others) but also the surface. There was another important development in 1976, when Wang *et al.* (1976) showed methane and nitrous oxide to be strong greenhouse gases as well. Both of these gases have natural sources, as well as anthropogenic ones (agriculture; natural gas; increase in cattle population etc). These two gases also interfered with the ozone chemistry and along with carbon monoxide (a pollutant) contributed to the increase in lower atmosphere ozone, which contributed to the surface warming (Fishman *et al.*, 1979). Until this study, lower atmosphere ozone was recognized only as a pollutant. Thus in a matter of five years after the discovery of the CFC greenhouse effect, chemistry emerged as a major climate forcing process (Figure 9). The global warming problem was not just a CO_2 problem but became recognized as a trace gas- climate change problem.

WMO's recognition and lead into IPCC

But it took five more years for the climate community to accept this view, when WMO commissioned a committee to look into the trace gas greenhouse effect issue. The report of this committee published as a WMO report in 1986 (Ramanathan *et al.*, 1986) concluded that trace gases other than CO_2 contributed as much as CO_2 to the anthropogenic climate forcing from pre-industrial times. This report also gave a definition for the now widely used term: Radiative Forcing, which is still used by the community. Shortly thereafter, in 1988, WMO and UNEP formed the Intergovernmental Committee Panel on Climate Change (IPCC), which in its 2001 report, confirmed that the CO_2 contributed about half of the total forcing and the balance is due to the increases in methane, nitrous oxide, halocarbons and ozone (Figure 10, see p. 243, from the 2001 report of IPCC). The anthropogenic radiative forcing from pre-industrial to now (year 2005) is about 3 Wm^{-2}, out of which 1.6 Wm^{-2} is due to the CO_2 increase and the balance is due to CFCs and other halocarbons, methane, nitrous oxide, ozone and others. The unit Wm^{-2} represents the number of watts added energy per square meter of the Earth's surface.

II. Predictions & Verifications of the Warming

When Will the Warming be Detected?

I now turn my attention to predictions of how the climate response would respond to the anthropogenic greenhouse forcing. As the trace gas importance began to emerge, I realized that the climate problem was a lot more serious than what we had thought; so I teamed up with the famous meteorologist Roland Madden and we started an analysis to see when were we going to detect this climate change and how much time did we have. Based on our analysis we made the prediction that if our greenhouse-global warming theory was reasonably accurate we should see the warming by the year 2000 (Madden and Ramanathan, 1980). This paper addressed some of the fundamental issues about natural climate variability. What we did was look at the temperature records of the North Atlantic and North Pacific, that's where we had a nearly homogeneous record of temperatures, and we quantified the climate noise, alternately the climate chaos, or natural variability. We quantified how that noise comes down with average in time (Figure 12, see page 243); i.e., the longer and the longer you average the noise comes down so that the signal (which should increase with time) can be detected. This is shown in Figure 11, which gives you an idea of the stochastic nature of detecting the climate change. Then we solved a couple of differential equations to look at what the expected temperature change was, and we also had to model how much of the emitted carbon dioxide would be airborne. Basically we concluded (quote from Madden and Ramanathan, 1980) '*Further consideration of the uncertainties in model predictions and of the likely delays introduced by ocean thermal inertia extends the range of time for the detection of warming, if it occurs to the year 2000*'. The IPCC report published in 2001 confirmed our predictions, when it concluded that the balance of evidence suggests a discernible human influence on climate. The observed surface temperature record (Figure 12) shows clearly how the temperature of the late twentieth century revealed the warming, although in 1980 (when we made our prediction), the warming was barely discernible.

I would also like to point out that the importance as well as the potential dangers of the global warming issue was well recognized more than 25 years ago by many scientists working in the field (e.g., Revelle; Manabe; Schneider; Hansen; Cicerone; Crutzen; Dickinson among others). For example, the Madden and Ramanathan (1980) paper began with the statement '*The possible climate effects of large increases in atmospheric CO_2*

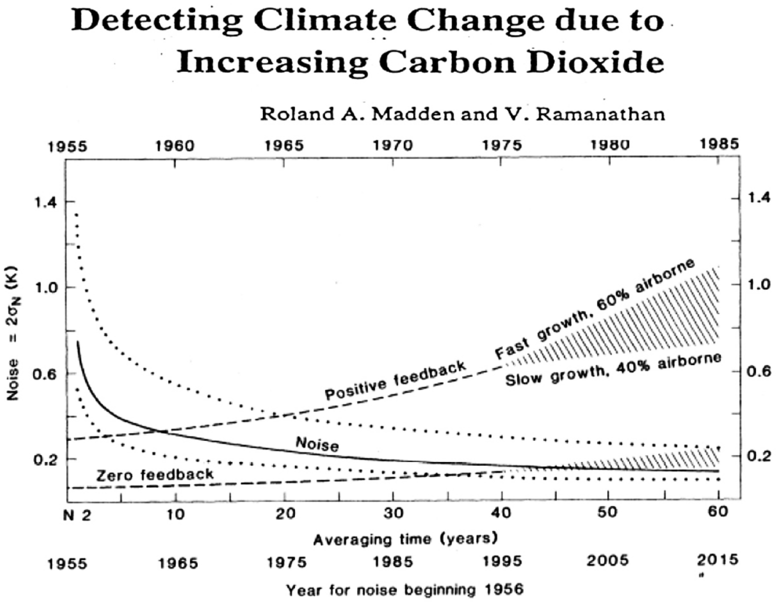


Figure 11. Madden and Ramanathan, 1980.

due to burning of fossil fuels may constitute one of the important environmental problems of the coming decades'. In a follow on paper written 8 years later (Ramanathan, 1988), I concluded that *'surface warming as large as that predicted by models would be unprecedented during an interglacial period such as the present'*.

A Tool for Verifying the Physics of the Warming

Thus far I deduced the greenhouse effect and the global warming using deductions from well understood physical laws. Now I would like to take up the issue of examining the magnitude of the forcing and the warming. This takes me back to the mid 1970s, when I was a post-doc at NASA Langley. My office happened to be in the same building where a group (headed by Dr G. Sweet, now deceased) was designing the Earth Radiation Budget Experiment (ERBE). I joined the ERBE science team and proposed to them that we should determine the atmospheric greenhouse effect quantitatively from the data. This was a satellite experiment that measured the incoming solar radiation, reflected solar radiation and the outgoing IR radiation (Figure 13 from Barkstrom, 1984, see page 244). Basically, what the satellite had was two scanning telescopes, which scanned the planet at 25-km footprint.

It took the team nearly 15 years from the concept stage to launch the satellite and publish the data! Using these data, we first quantified (Ramanathan *et al.*, 1989) that the incoming sunlight was 342 Wm^{-2}, out of which 102 Wm^{-2} was reflected to space; and the IR energy (heat radiation) leaving the planet was 237 Wm^{-2} (Figure 14, see page 244). The uncertainty in these estimates was about 5 Wm^{-2}. Let me explain the units. The W stands for the unit of energy flow per second which is Watts. The 'm^{-2}' denotes the energy flow per square meter of the Earth's surface. ERBE offered a powerful tool for verifying some of the scientific underpinnings of the theory of greenhouse effect and global warming.

Natural Greenhouse Effect: How Large is it?

We are now ready to put a number to the thickness of Nature's blanket. To determine the greenhouse effect of the planet, we estimated the IR energy given off by the surface. All this requires is data on the surface temperature (Ts) of the planet and the so-called emissivity of the surface, which is close to 1 for sea surface and varies by about 0.85 to 1 for land surfaces depending on the soil moisture. The IR energy emitted by the surface is about 399 Wm^{-2} (Figure 15, see page 245). The next quantity we need to know is the IR energy emitted by the atmosphere. To estimate this, we scanned the telescope in between clouds to extract the emission from clear skies (Ramanathan *et al.*, 1989) and the annual and global average of the emission was 268 Wm^{-2}. So the planetary surface emits 399 Wm^{-2} out of which only 268 Wm^{-2} escapes to space and the difference of 131 Wm^{-2} is trapped in the atmosphere, which is the greenhouse effect of the planet during 1985 to 1989. We have to subtract the anthropogenic greenhouse effect of a few Wm^{-2} to get the natural value. By comparing with the absorbed solar radiation (341-102=239 Wm^{-2}), it is clear the natural greenhouse effect is very large, without which the planet would be cold and frozen.

A Metric for Judging the Anthropogenic Greenhouse Effect

We now have a metric for assessing the manmade greenhouse effect. We have added 3 Wm^{-2} to the greenhouse effect by adding CO_2 and other greenhouse gases to the atmosphere. Comparing this with the observed greenhouse effect of 131 Wm^{-2}, we note that human activities have basically thickened the blanket by about 2.5%.

Magnitude of the Warming: Amplification by Water Vapour Feedback

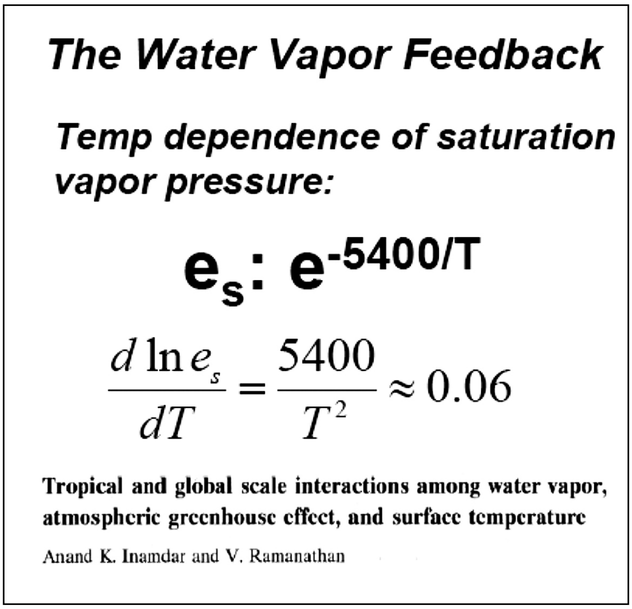
The next issue is, how large is the warming going to be, given the forcing of 3 Wm^{-2}. Of course, we can run climate models to estimate this, but I was interested in obtaining an independent estimate solely from observations. I started working on this problem from the early 1990s and obtained the estimate by the late 1990s (Inamdar and Ramanathan, 1998).

Recall that I said, the planet will warm until it radiates the excess IR energy (3 Wm^{-2} in our case) to space. All we need to know is the rate at which the surface and the atmosphere radiate energy per degree warming… a fundamental number for the planet. Before we address this, we need to deal with one important complication concerned with water vapour thermodynamics.

Thermodynamics of water vapour, as given by the Clausius-Clapeyron equation (Figure 16) dictates that the saturation vapour pressure, es, of water vapor increases exponentially with temperature. If you take the differential of that equation with temperature, you find that at a temperature of 300K (close to surface temperature of the tropical oceans) the vapour pressure increases by about 6% for each 1 degree rise in temperature. The percent increase is more at the colder atmospheric temperatures (7% to 10%). In essence, as the atmosphere warms, its moisture holding capacity increases. This is why, for example, in the extra tropics the summer is humid and the winter is dry. When it gets cold it gets dry because the vapour pressure drops precipitously. But the issue for us is that the water vapour is the strongest greenhouse gas in the planet.

Let's do a simple thought experiment. Before the industrial era, the planet is in energy balance with incoming sunlight. With the dawn of industrialization, we are adding CO_2 and other greenhouse gases and as a result the outgoing heat (IR) radiation decreases.

There is excess energy and the surface and the atmosphere begins to warm. But, as the atmosphere warms, the water vapour concentration begins to increase at the rate of 6% to 10% per degree warming; since water vapour is a strong greenhouse gas, this increase will amplify the greenhouse warming. Arrhenius was well aware of this feedback effect and included it. So when we attempt to estimate the rate of emission per degree warming, we have to account for this feedback.



The Water Vapor Feedback

Temp dependence of saturation vapor pressure:

$$e_s : e^{-5400/T}$$

$$\frac{d \ln e_s}{dT} = \frac{5400}{T^2} \approx 0.06$$

Tropical and global scale interactions among water vapor, atmospheric greenhouse effect, and surface temperature

Anand K. Inamdar and V. Ramanathan

Figure 16.

Fortunately the planet does a major geophysical experiment every year. When you take the entire planet, its mean surface temperature in June-July-August is about 4 degrees Celsius warmer than during Dec-Jan-Feb. We had the ERBE and other observations at 200 km spatial scale for 1985 to 1989, and so we used it to estimate the greenhouse effect, Ga, the water vapour amount in three layers using a microwave instrument on another satellite and surface temperature data. These are shown in Figure 17 as a function of month. What we see is that, as surface temperature increases from Jan to July, water vapour amount in the three layers (between surface to 12 km) increases by about 7% per degree Celsius increase; and the atmospheric greenhouse effect increases at the rate of 3.5 Wm^{-2} per degree increase in surface temperature. Now we are ready to get the fundamental number we want. As shown in the bottom right panel, the outgoing heat radiation given off by the planet increases by 2.1 Wm^{-2} C^{-1}, i.e., the surface-atmosphere system gives off 2.1 Wm^{-2} of energy per Celsius increase in its temperature. I will subtract about 0.2 for solar absorption by the increase in water vapour (a minor detail) to get the final number of 1.9 Wm^{-2} C^{-1}

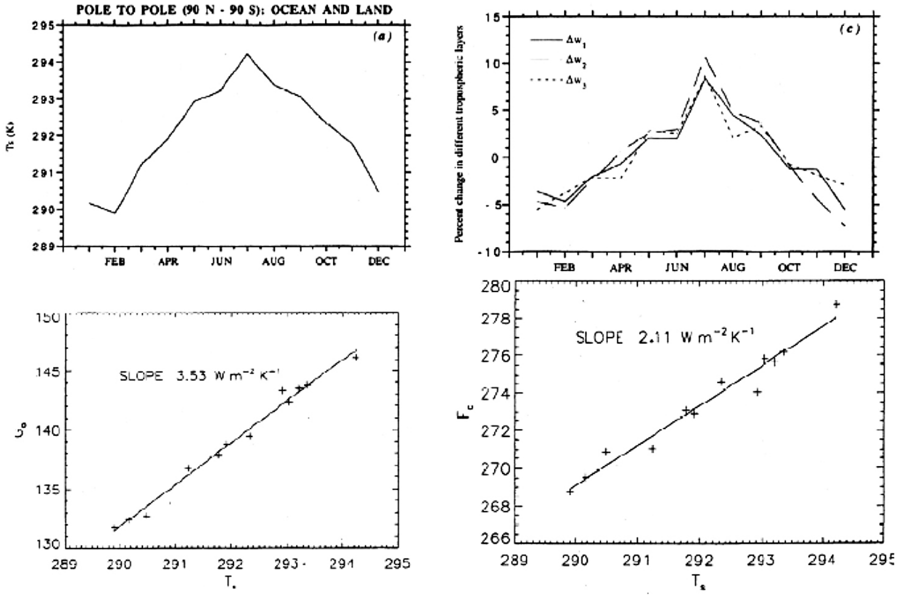


Figure 17.

(Figure 18, see page 245). These are all remarkably close to what we would expect from thermodynamics and basic radiation physics. Without the water vapour feedback, this number would have been about 3.4 Wm^{-2} C^{-1} instead of 1.9 Wm^{-2} C^{-1}, i.e. the planet would have gotten rid of more energy with warming without the thermodynamic coupling of water vapour with surface temperature. Alternately, the climate system is less stable with the positive water vapour feedback.

Warming Commitment: 2°C

Now we have all of the basic elements to put the pieces together. Since the dawn of the industrial era (say 1850) to now, the added greenhouse gases have trapped 3 Wm^{-2} of energy to the planet. It begins to warm. The planet can get rid of only 1.9 Wm^{-2} per degree Celsius increase in temperature. So it needs to warm by (3/1.9=) 1.5 °C (Figure 18) to restore the energy balance. A warming of this large magnitude, will melt and retreat sea ice and snow packs in mountain glaciers. Such melting will expose the underlying darker surfaces (ocean and rocks) which will absorb even more sunlight and thus amplify the warming. As the planet is warming and giving off 1.9 Wm^{-2} per degree warming, it begins to absorb about 0.4 Wm^{-2} more sunlight (this number is from models and hence is a bit soft) and thus it can get rid of only 1.5 Wm^{-2} $°C^{-1}$. Thus the required warming to balance the 3 Wm^{-2} greenhouse forcing, is (3/1.5) 2 °C. In effect, we have committed the planet to a 2 °C warming. I had used this term 'commitment' in a paper written in 1988 (Ramanathan, 1988) and it is nice to see its prevalent use now.

The Missing Warming of About 1 °C

A closer scrutiny of the temperature record (Fig. 12, see page 243) reveals that the observed warming is only about 0.7 °C, compared with our prediction of 2 °C. What happened to the rest? One obvious explanation is our model of climate sensitivity (the 1.5 Wm^{-2} $°C^{-1}$ number) is wrong. But let us look closer at the details… after all the devil is in the details. First, the ocean has a huge thermal inertia. It mixes the heat by turbulence quickly (within weeks to months) to the first 50 to 100 m depth. From there, the large scale ocean circulation mixes the heat in about few years to few decades to about 500m to 1000m depth. Some of the excess energy trapped is still circulating in the ocean. Oceanographers have estimated that about 0.6 (0.2) Wm^{-2} of the 3 Wm^{-2} is still stored in the ocean. So about 0.4 °C of the warming will show up in the next few decades. Thus we have accounted for 1.1 °C (0.7+0.4) of the 2 °C committed warming. What happened to the remaining 0.9 °C? Although my prediction with Dr Madden (that the warming would be detected by 2000) was verified, I could take little comfort in that because the magnitude of the warming was almost a factor of two less than what we had predicted.

So the search was on for the missing 0.9 °C warming which led me to the masking effect of Atmospheric Brown Clouds (ABCs). But the path to ABCs was not a straight forward one, and I did not realize at that time it would take me more than 15 years to track the 0.9 °C down, but it was en exciting detour.

The Super Greenhouse Effect and a Thermostat in the Pacific Warm Pool

By the late 1980s it became clear to me that we have to look for analogues for a warmer planet. This quest took me to the western Pacific warm pool, the largest body of the warmest ocean in the planet (Figure 19, see page 246), somewhat accidentally. An undergraduate student, A. Raval, joined my lab to do research. I normally do not take undergraduate students for research, but Raval was recommended by a Nobel laureate in Physics at Univ of Chicago. We (Raval and Ramanathan, 1990) started using ERBE data for the greenhouse effect (Ga) to see how it increased with temperature due to the water vapour feedback (Figure 20, see page 246). We focused only over the oceans, because the ERBE data for land was not as accurate. Up to about a temperature of 20 °C (293 K in the figure) the Ga increased with sea temperature as expected from water vapour thermodynamics; but it started increasing more rapidly beyond 20 °C and for sea surface temperatures warmer than 27 °C, it increased at an unstable rate, referred to as super greenhouse effect. While the ERBE curve was published in 1990 (Raval and Ramanathan, 1989), the NCAR CCM3 with coupled ocean model simulated very warm climates similar to that experienced in the Cretaceous and the model was able to simulate the super greenhouse effect (see the red dots in Fig. 20). In the observations, the super greenhouse effect originated in the warm pool since this is the region where surface temperatures exceeded 27 °C. The question raised by this behaviour was: What process is preventing the western Pacific Warm pool temperatures to increase unstably? If any, it is a remarkable fact that the seasonally average warm pool surface temperatures rarely exceed 31 °C! As I was wondering about this remarkable situation, Raval's friend, W.D. Collins, an astrophysics grad student from Chicago, heard about this super greenhouse effect and expressed interest in joining my lab after his Ph D. He was recommended strongly to me by another Nobel laureate, Prof. S. Chandrasekhar (my neighbour and friend at U of Chicago), and Bill Collins joined my lab in early 1990, as I was getting ready to move to the Scripps Institution of Oceanography.

Bill and I started working on the problem and, after a year of intense study extending well into the night many days, came up with the thermostat hypothesis (Ramanathan and Collins, 1991). We proposed that, as the sea surface warms in response to the intense tropical solar radiation and the super greenhouse effects, it begins to form deep convective and thick cirrus anvil clouds. The widespread convective anvils reflect solar radiation and shield the sea surface from the intense solar radiation. In support of this hypothesis, we pointed out that the warm pool was very humid but mostly cloudy. In addition, we pointed out that even during an El Niño, when the normally cold central Pacific warms intensely, its temperature do not exceed 31 °C and the region which in normal years is mostly free of anvil clouds is filled with convective anvils during an El Niño. This hypothesis was strongly contested and we mounted a field experiment in 1993 to test it, thanks to the keen support of Dr Jay Fein (director of climate research at NSF) and the mentoring by Dr J. Kuettner, a legendary scientist, glider pilot and a pioneer in conducting complex field campaigns.

The Central Equatorial Pacific Experiment (CEPEX; Fig 21, see page 247), was my first entry into the world of field experiments, which became a life long passion. Teaming up with Dr Kuettner (who was 82 years old then), we deployed aircraft, ships and satellites (in search of the thermostat) to study the heat budget of the warm pool. There was another major surprise waiting for us in the warm pool. While the surface and the aircraft data confirmed our satellite findings of the warm pool to be very humid, very cloudy and subject to the super greenhouse effect, the sunlight reaching the warm pool surface was substantially (by about 8%) less than the value predicted by the best models we had (Ramanathan *et al.*, 1995). We had constrained the model at the top of the atmosphere by satellite data and thus the discrepancy could not be because the model clouds were not reflecting to space the correct amount. The discrepancy had to be due to the fact that our model atmosphere was not absorbing enough solar energy, i.e., it was missing an important process that was absorbing solar radiation in the atmosphere. I had to postpone my search for the thermostat and solve this puzzle first. This quest took me to the issue of aerosols, i.e., particles in the atmosphere. I had totally ignored this topic, although it had been pursued actively by several scientists since the early 1970s (e.g., see Rasool and Schneider, 1973). I knew I had to account for atmospheric aerosols in the model, particularly black carbon in soot, which was a major absorbing species. This enquiry, ultimately, led to new insights into the missing surface warming of 0.9 °C. But many more dots had to be connected first.

III. Masking of Warming by Atmospheric Brown Clouds

Synopsis

In addition to adding greenhouse gases, human activities also contributed to the addition of aerosols (sub micron size condensed particles) to the atmosphere. Since 1970 (Mitchell, 1970) scientists have speculated that these aerosols are reflecting sunlight before it reaches the surface and thus contribute to a cooling of the surface. This was further refined by Charlson *et al.* (1990) with a chemical transport model. They made an estimate of the cooling effect of sulfate aerosols (resulting from SO_2 emission) and concluded that the sulfate cooling may be substantial. Essentially, aerosols concentrations increased in time along with greenhouse gases, and the cooling effect of the aerosols has masked some the greenhouse warming. I am choosing the word 'mask' deliberately (Figure 22, see page 247), for when we get rid of the air pollution, the masking will disappear and the full extent of the committed warming of 2 °C will show up. Several tens of groups around the world are working on this masking effect using models and satellite data, but my route was a bit more tortuous and adventurous.

Indian Ocean Experiment: In Search of the Missing Absorption

In search of the missing solar absorption identified by CEPEX, I decided the Arabian Sea would be the ideal place for it, in part, because it was also a very warm ocean and very humid, not unlike the warm pool. Furthermore, during the 6 month long dry season, air from S Asia laden with pollution was blowing over the Arabian Sea on its way towards the inter tropical convergence zone south of equator. My colleague, Dr Paul Crutzen, became interested in this idea, because of his longstanding interest in air pollution in the tropics. Over a brief lunch at Scripps in 1994, Paul and I decided to look at how transport of air pollution from S Asia would impact the Indian Ocean. However it was difficult to get the funding based solely on this. We decided to broaden the scope to quantify the cooling effect of the aerosols… basically quantify the missing warming, i.e., the thickness of the mask. What started out as a simple experiment with one ship, emerged into a major international field experiment that would ultimately cost $25 M, the Indian Ocean Experiment (INDOEX) with participants from India, Maldives, Europe and USA (Figure 23 from Ramanathan *et al.*, 1996 in http://www-ramanathan.ucsd.edu/field_exp.html, see page 248). Again Dr Jay Fein of NSF played a major role in making this happen, my colleague Mr Hung Nguyen was a critical player in the execution, and Dr A.P. Mitra of India successfully mounted a major effort by Indian scientists to play a major role in INDOEX. We started ship observations in 1997, set up an observatory in the Maldives in 1997 and conducted a major field campaign in 1999 from the Maldives with 6 aircraft, 2 ships and surface observatories with over 200 scientists from Europe, India and USA, with Paul and me as the Co-Chief scientists. I was responsible for the field campaign in Maldives and it was a remarkable experience to lead this once in a lifetime campaign with impressive scientists from around the world speaking over 12 languages.

Widespread Brown Clouds

Brown clouds are usually associated with the brownish urban haze such as the one shown in the photograph (Figure 24, see page 248) taken by me flying over Los Angeles (Dec. 27, 2002). What we discovered instead during INDOEX was widespread brownish haze over most of S Asia flowing into the Indian Ocean (Figure 25, see page 249). This was due to fast long range transport by winds. Flying on a C-130 loaded with instruments, it became quickly evident that we were dealing with a huge problem. After the completion of the field campaign, NASA released a new aerosol instrument (MODIS) on the TERRA satellite and we analyzed the data which revealed that the brown clouds were not just an S Asian problem, but a worldwide issue (Figure 26, see page 249). As described next, the brownish color was due to strong solar absorption by black carbon in the soot.

Fingerprinting the Source of Missing Solar Absorption

Filters collected from the aircraft were analyzed by transmission electron microscope (by Dr J. Anderson) which revealed how the soot particles were attached to other aerosols and traveled as far south as 6S into the southern Indian Ocean (Figure 27, see page 250). It was only after we crossed the inter tropical convergence zone, south of 8S (close to Diego Garcia), that we were rid of the brown clouds. As mission scientist on one of the flights, I distinctly recall requesting the pilot to go as far south as possible until we see beautiful clear skies. The pilot informed me that he could not fly no further south, for we were about 1500 km from home and the aircraft was close to its endurance limit. Chemical analysis revealed that the brown cloud aerosols consisted of strongly absorbing black carbon and in addition, sulfates, nitrates and organics which were reflecting solar radiation and thus were the masking agents of global warming (Figure 28, see page 250). So, we had found both the source for both the missing absorption and the masking of global warming. What remained to be done was to quantify the energy absorbed and energy reflected. We had radiometers on the surface, ships, satellites and aircraft for this purpose.

Measuring the Solar Energy Absorbed and Reflected

We had deployed grating spectrometer to measure high resolution solar spectrum and as the ship traveled in and out of the plume, my post doctoral fellow, Dr Jens Meywerk, would take a spectrum of the direct sunlight and the reflected (downwards) solar radiation. The data revealed (Figure 29, see page 251) that the brown clouds led to a large reduction in sunlight, with the largest reduction of 40% in visible wavelengths (another indication of soot absorption); in addition, the data also quantified the reflected solar radiation, also shown in Figure 30 (see page 251). We needed one more piece of data before estimating the energy absorbed and reflected. This dealt with how the particles interacted with clouds. This issue arises because cloud drops are nucleated by aerosols and the manmade aerosols such as sulfates are very efficient in nucleating. Drs Heymsfield and McFarquhar, took measurements from C0130 and demonstrated that clouds embedded in pollution had an order of magnitude more cloud drops. Thus the polluted clouds with more drops will scatter more sunlight and lead to more cooling. This data plotted with other data worldwide (Figure 31, see page 252) shows how aerosols in brown clouds lead to increased cloud drops worldwide.

Finally, the energy absorbed and reflected to space by the ABCs (direct forcing in Figure 31) and by the ABC's influence on clouds (indirect forcing in Fig. 31) over the Indian Ocean is shown along with the radiative forcing due to greenhouse gases. Soot in the brown clouds increased solar absorption by about 14 Wm^{-2}, which is as much as 20% of the 70 Wm^{-2} absorbed by the background atmosphere. It also shows that at the top of the atmosphere, ABCs by reflecting solar radiation back to space, have reduced the net solar energy coming into the system by 5 Wm^{-2}. This is part of the masking effect of warming we were looking for, but we need a global average estimate. In short, INDOEX helped find the missing solar absorption; but the data in Fig. 31 raised another issue that caught me by surprise.

Dimming

Figure 31 (see page 252) also shows that ABCs reduce the solar radiation reaching the surface by as much as 20 Wm^{-2}, which is as much as 10% of the solar radiation absorbed by the Indian Ocean. In effect the surface is dimmer by about 10% during 1999, due to the shielding of surface by ABCs aloft. This raised two key questions: *How long has the dimming been going on? And what is its implication to regional climate.*

In order to examine the first question, we modeled the historical variations in ABCs and their dimming influence, by including historical variations in emissions of soot and SO_2 in the NCAR climate model which I described in the beginning. Fortunately, we had well calibrated solar radiation data over India (12 stations) that was collected by a well-known Indian meteorologist, Dr Annamani. She was a dedicated scientist intent on the accuracy of the data. The results (Ramanathan *et al.*, 2005) are shown in Fig. 32 (see page 252) along with the simulated values. First the observations reveal that India has steadily been getting dimmer at least from the 1960s (data record began in the 1960s) and that India now is about 7% dimmer than in the 1960s. Next, the simulations were able to track down observations reasonably well and they attributed the cause to the 4 to 5 fold increase in emissions of soot and SO_2. We will take up the next question next.

Impact on the Monsoon and Rice Harvest

Solar radiation at the surface is the fundamental source for evaporation of moisture. Hence if we reduce solar radiation, it is likely we will reduce evaporation and in turn the rainfall. As a result the most direct result of dimming is to reduce rainfall. But where? To answer this we turned to the NCAR climate model and the simulations suggested that the observed reduction in monsoon rainfall during the last 50 years is most likely due to the ABCs induced dimming (Figure 33, see page 253). The simulations revealed another way in which ABCs were slowing down the monsoon circulation. Since the ABCs were concentrated in the Northern Indian Ocean and S Asia, their cooling effects were concentrated there which reduced the north to south gradient in sea surface temperatures, which was also responsible for the rainfall decrease.

This result made me curious about the impact of this dimming and the long term rainfall decrease on food security of India. I teamed up with two agricultural economists from UC San Diego (Dr J. Vincent) and UC Berkeley (Dr M. Aufhammer) to model the impact on rain fed rice harvest. The integrated agro-climate model, a statistical model, suggested that ABCs have led to a 11% reduction in rice harvest, while the surface warming (due to greenhouse gases) has led to a 4% reduction in the harvest. Is there evidence for such a decrease in the actual harvest? It turns out, while rice harvest increased rapidly in the 60s due to green revolution, the harvest leveled off by the 1990s (Figure 34, see page 253).

Project ABC: The Next Step

In short, INDOEX data and the follow on modeling work helped identify ABCs as a major issue threatening the water and food security of India. This was just a beginning. The MODIS satellite data, that was analyzed after the INDOEX experiment (Figure 10, see page 243), revealed that we sought UNEP's help in organizing the project. All it took was one flight over the Nepal-Himalayas with Dr K. Toepfer, the head of UNEP in 2002 and Dr S. Shrestha, the head of UNEP's Asia office. They immediately formed the ABC project and brought in a team of scientists from Asia. We began setting ABC observatories (Figure 35, see page 254) and integrated the field data with satellite observations and aerosol-transport models to determine the dimming and solar absorption by ABCs over Asia and rest of the world (Chung *et al.*, 2005; Ramanathan *et al.*, 2007). Finally after nearly 15 years of detour through thermostat, CEPEX and INDOEX, I had a global view of the missing warming.

Magnitude of the Missing Warming

Global distribution of atmospheric solar heating and surface dimming by ABCs for the 2001-2003 period is shown in Figure 36 (see page 254). Both reveal peak values over polluted regions in S and E Asia, Eastern N America, Amazon, Southern Africa, Indonesia, etc. Focusing over Asia, the strong soot induced heating surrounding the Himalayan-Hindu-Kush region is contributing to the retreat of these glaciers and snow packs, a dominant source of many major river systems in the region. The dimming is spreading over the Indian Ocean, tropical Atlantic ocean and the western Pacific ocean, with implications to regional water budget over many tropical nations. The global average of the ABC forcing is compared with the greenhouse forcing in Fig. 37 (see page 255). Focusing just on the top of the atmosphere forcing, ABCs have led to a negative forcing of -1.4 Wm^{-2} (add the direct and the indirect panels), i.e., have reflected back to space 1.4 Wm^{-2} of the incoming solar energy. Comparing this with the 3 Wm^{-2} greenhouse forcing, we see that ABCs have masked about 50% of the warming.

IV. Uncertainties

Will Clouds Rescue us from Severe Warming or Make it Worse?

What's going to happen to the clouds in a warmer world? Will a warmer planet become cloudier? Why is this even an issue? The topic that addresses these questions is known as cloud feedback. It is a topic that worries me the most about the future. I started to work on it during the mid to 1980s but had to put it aside due to the detours I discussed earlier. We know clouds look white because they reflect a lot of sunlight back to space; and clouds also have an enormous infrared greenhouse effect because they are, after all, water molecules. Models (Manabe and Wetherald, 1967; Schneider, 1972) were suggesting that the cooling effect was larger than the warming effect and clouds had a net cooling effect. But observational determination of the effect became a passion for me.

I must now return to the ERBE satellite data to complete this story. I persuaded NASA to sort out the radiation budget over clear skies (i.e. pull the data for the gaps in between clouds). This is referred to as clear sky radiation budget. When we subtract the all-sky values (clear plus cloudy, which is what the satellite normally sees) from the clear sky values, we get an estimate of how clouds are regulating the radiative heating of the planet and this is referred to as cloud-radiative forcing (Charlock and Ramanathan, 1983).

This data analysis confirmed earlier model simulations that clouds have a major global cooling effect on the planet (Ramanathan *et al.*, 1989); i.e., they were significantly reflecting more sunlight (top panel in Figure 38, see page 255) than absorbing the IR. What was interesting about this experiment was the new questions it raised about the cloud feedback issue:

> In the tropics the greenhouse effect and the solar effect were large but they nearly cancelled each other. How will this delicate balance be perturbed by a large warming? Will the greenhouse effect become larger (e.g., due to clouds reaching higher in a warmer planet) and amplify the warming?

The global cooling effect was due to the extra tropical, storm track cloud systems. I know these systems are the source for huge weather problems for US and Europe but at least we can take comfort in the fact that these systems are keeping the planet cooler. Basically, what this experiment showed was that clouds were acting like two giant umbrellas centred over the Artic and the Antarctic and shielding the planet from solar radiation. The magnitude of this shielding effect is so large that, if these clouds were to expand just by 1.5%, that's enough to compensate all of the 3 Wm^{-2}; or if they shrink in response to global warming, they can amplify it by a factor of two or more. The shrinking scenario is not unlikely, because the arctic sea ice is retreating rapidly in response to global warming and this is amplifying the arctic warming. The resulting decrease in equator to polar temperature gradients, can lead to a pole ward retreat of the storm track cloud systems.

How will the Biosphere Respond?

The second major surprise in store for us in the coming decades may very well be the biosphere response. You can ask, why was this not an issue in the last 100 years; and why is it an issue for the future? If the planet warms at the rate it has been warming for the next 50 years, then the climate will have gone beyond any stage we can go back in the past and model. It is well-known that the greenhouse gases like CO_2, CH_4 and N_2O are regulated by the biosphere (Figure 39, see page 256). But cloud formation is also influenced by the biosphere. For example, dimethyl-sulphide emitted by planktons in the ocean is the precursor for sulphates, efficient cloud condensation nuclei in the marine atmosphere. How will climate change perturb the regulation of greenhouse gases and clouds by biota (e.g. see Charlson *et al.*, 1987)?

V. Mitigating Unprecedented Climate Changes

I want to end on a positive note. At the rate of current increase in energy combustion and atmospheric CO_2 increase, it is likely that we are on a path to a future climate that is about 2 to 4 °C warmer in this century. For example, just by cleaning up the atmosphere of all aerosols in ABCs, we will remove the masking effect and contribute to additional warming of 1 to 1.5 °C. I am not implying that we should keep the pollutants in the air, for they cause serious health impacts and ecosystem impacts in the form of acid rain etc. But, the need to clean up the air increases the urgency for decreasing the growth of CO_2 in the air. We have to urgently seek and implement solutions for mitigating the climate change, and there are many that we can implement immediately.

The ABC research also offers hope for mitigating ABC effects on global warming and HHK glacier retreat. It has identified soot as the major villain in the negative effects of ABCs. Fortunately, we have the technology and the financial resources to significantly reduce soot emissions. Cooking with wood, coal and cow dung fires is the major source for soot emissions in many parts of S Asia and East Asia. Replacing such solid fuel cooking with solar and biogas plants is an attractive alternative. The lifetime of soot is less than a few weeks and as a result the effect of the deployment of the cleaner cookers on the environment will be felt immediately. To understand the socio-economic-technology challenges in changing the cooking habits of a vast population (700 million in India alone), we have started Project Surya with engineers, social scientists and NGOs in India (Figure 40, see page 256). For its pilot phase, Surya will adopt two rural areas: one in the HHK and the other in the Indo-Gangetic plains with a population of about 15000 each and deploy locally-made solar cookers and biogas plants. The unique feature is the Surya will accurately document the positive impacts of soot elimination on human health, deposition of soot on the glaciers, atmospheric heating and surface dimming.

Additional details of Surya can be found in (Ramanathan and Balakrishnan, 2006. http://www-ramanathan.ucsd.edu/ProjectSurya.html)

By improving the living conditions of the rural poor (average earning is less than 2$ a day) and by minimizing the negative health impacts of indoor smoke, Surya is a win-win proposition. Surya is but one example of how each one of us must think of practical and innovative ways for solving the global warming problem. Science has provided us with immense knowledge of the impact of humans on the climate system and we have to use this knowledge to develop practical solutions that combine behavioral changes with adaptation and mitigation steps.

Acknowledgments

I thank P.J. Crutzen, J. Fein, J. Kuettner and H. Nguyen for their support and friendship. I thank Giri Ramanathan for her insightful suggestions on how to organize this lecture. I thank the NSF, NASA and NOAA for supporting this research.

REFERENCES

Arrhenius, S. (1896). On the influence of carbonic acid in the air upon the temperature of the ground, *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science, 41*, 237-276.

Auffhammer, M., V. Ramanathan, and J.R. Vincent (2006). Integrated model shows that atmospheric brown clouds and greenhouse gases have reduced rice harvests in India, *PNAS*, 10.1073/pnas.0609584104.

Augustsson, T. and V. Ramanathan (1977). A Radiative-Convective Model Study of the CO_2-Climate Problem, *J. Atmos. Sci., 34,* 448-451.

Barkstrom, B.R. (1984). The Earth Radiation Budget Experiment. *Bulletin of the American Meteorological Society*, 65, 1170-1185.

Charlson, R.J., J.E. Lovelock, M.O. Andreae, and S.G. Warren (1987). Oceanic phytoplankton, atmospheric sulfur, cloud albedo and climate, *Nature, 326*, 655-661.

Charlson, R.J., J. Langner and H. Rodhe (1990). Sulphate aerosol and climate, *Nature, 348*, 22.

Chung, C.E., V. Ramanathan, D. Kim and I.A. Podgorny (2005). Global anthropogenic aerosol direct forcing derived from satellite and ground-based observations, *J. Geophys. Res., 110*, D24207, doi:10.1029/2005JD006356.

Crutzen, P.J. (1972). SSTs: a threat to the earth's ozone shield, *Ambio, 1*, 41-51.

Donner, L. and V. Ramanathan (1980). Methane and Nitrous Oxide: Their Effects on the Terrestrial Climate. *J. Atmos. Sci., 37,* 119-124.

Fishman, J., V. Ramanathan, P.J. Crutzen and S.C. Liu (1980). Tropospheric Ozone and Climate. *Nature, 282,* 818-820.

Inamdar, A.K. and V. Ramanathan (1997). On Monitoring the Atmospheric Greenhouse Effect from Space, *Tellus, 49B,* 216-230.

Inamdar, A.K. and V. Ramanathan (1998). Tropical and Global Scale Interactions Among Water Vapor, Atmospheric Greenhouse Effect, and Surface Temperature. *J. Geophys. Res. 103*, D24: 32177-32194.

IPCC, (2007). *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, [Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 996pp.

Le Treut, H. and R. Somerville, U. Cubasch, Y. Ding, C. Mauritzen, A. Moksslt, T. Peterson and M. Prather, (2007). 'Historical Overview of Climate Change', In: *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, [Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Lorenz, E.N. (1972). Barotropic instability of Rossby wave motion, *J. Atmos. Sci., 29*, 258-264.

Madden, R.A. and V. Ramanathan (1980). Detecting Climate Change due to Increasing Carbon Dioxide, *Science, 209,* 763-768.

Manabe, S. and R.T. Wetherald (1967). Thermal equilibrium of the atmosphere with a given distribution of relative humidity, *J. Atmos. Sci., 24,* 241-259.

Molina, M.J. and F.S. Rowland (1974). Stratospheric sink for chloroflurormethanes: chlorine atomc-atalysed destruction of ozone, *Nature, 249,* 810-812.

Pitcher, E.G., R.C. Malone, V. Ramanathan, M.L. Blackmon, K. Pure and W. Bourke (1983). January and July Simulations with a Spectral General Circulation Model, *J. Atmos. Sci., 40,* 580-604.

Ramanathan, V. (1975). Greenhouse Effect Due to Chlorofluorocarbons: Climatic Implications. *Science, 190,* 50-52.

Ramanathan, V., (1980) 'Climatic Effects of Anthropogenic Trace Gases', in *Interactions of Energy and Climate*, W. Bach, T. Pankrath and J. Williams, (eds.) D. Reidel Publishing Co., pp. 269-280.

Ramanathan, V. (1988). The Greenhouse Theory of Climate Change: A Test by an Inadvertent Global Experiment, *Science, 240,* 293-299.

Ramanathan, V. and A. Inamdar, (2006) 'The Radiative Forcing due to Clouds and Water Vapor' in *Frontiers of Climate Modeling*, J. T. Kiehl and V. Ramanthan, (eds.), Cambridge University Press, pp. 119-151.

Ramanathan, V. and K. Balakrishnan (2006). Project Surya: reduction of air pollution and global warming by cooking with renewable resources, a white paper, 14pp.

Ramanathan, V. and R.D. Cess (1974). Radiative transfer within the mesospheres of Venus and Mars, *The Astrophyical Journal, 188,* 407-416.

Ramanathan, V. and M.V. Ramana (2005). Persistent, Widespread, and Strongly Absorbing Haze Over the Himalayan Foothills and the Indo-Ganges Plains, *Pure and Applied Geophysics, 162,* 1609-1626.

Ramanathan, V. and W. Collins (1991). Thermodynamic Regulation of Ocean Warming by Cirrus Clouds Deduced from Observations of the 1987 El Niño. *Nature, 351,* 27-32.

Ramanathan, V., B.R. Barkstrom, and E.F. Harrison (1989). Climate and the Earth's Radiation Budget. *Physics Today, 42(5),* 22-33.

Ramanathan, V., B. Subasilar, G. Zhang, W. Conant, R. Cess, J. Kiehl, H. Grass and L. Shi, (1995). Warm Pool Heat Budget and Shortwave Cloud Forcing: A Missing Physics? *Science, 267,* 499-503.

Ramanathan, V., C. Chung, D. Kim, T. Bettge, L. Buja, J.T. Kiehl, W.M. Washington, Q. Fu, D.R. Sikka, and M. Wild (2005). Atmospheric Brown Clouds: Impacts on South Asian Climate and Hydrological Cycle. *PNAS, 102,* No. 15, 5326-5333.

Ramanathan, V., E.J. Pitcher, R.C. Malone and M.L. Blackmon (1983). The Response of a Spectral General Circulation Model to Refinements in Radiative Processes. *J. Atmos. Sci., 40,* 605-630.

Ramanathan, V., *et al.*, (2001). The Indian Ocean Experiment: An Integrated Assessment of the Climate Forcing and Effects of the Great Indo-Asian Haze. *J. Geophys. Res., 106*, D22, 28371-28399.

Ramanathan, V., F. Li, M.V. Ramana, P.S. Praveen, D. Kim, C.E. Corrigan, H. Nguyen (2007). Atmospheric Brown Clouds: Hemispherical and regional variations in long range transport, absorption, and radiative forcing. *J. Geophys. Res., 112*, D22S21, doi:10.1029/2006JD008124.

Ramanathan,V., L.B. Callis and R. E. Boughner (1976). Sensitivity of Surface Temperature and Atmospheric Temperature to Perturbations in Stratospheric Concentration of Ozone and Nitrogen Dioxide. *J. Atmos. Sci., 33,* 1092-1112.

Ramanathan, V., P.J. Crutzen, J.T. Kiehl and D. Rosenfeld, (2001). Aerosols, Climate, and The Hydrological Cycle, *Science, 294*, 2119-2124.

Ramanathan, V., R.D. Cess, E.F. Harrison, P. Minnis, B.R. Barkstrom, E. Ahmad, and D. Hartmann (1989). Cloud-Radiative Forcing and Climate: Results from the Earth Radiation Budget Experiment, *Science, 243*, 57-63.

Ramanathan, V., R.J. Cicerone, H.B. Singh and J.T. Kiehl (1985). Trace Gas Trends and Their Potential Role in Climate Change. *J. Geophys. Res., 90,* 5547-5566.

Raval, A. and V. Ramanathan (1989). Observational Determination of the Greenhouse Effect. *Nature, 342*, 758-761.

Rodhe, H., R. Charlson and E. Crawford (1998). 'Svante Arrhenius and the Greenhouse Effect' in: *The Legacy of Svante Arrhenius Understanding the Greenhouse Effect*, H. Rodhe and R. Charlson (eds.), Royal Swedish Academy of Sciences, Stockholm, pp. 13-20.

UNCERTAINTIES IN CLIMATE CHANGE SCIENCE

MARIO J. MOLINA

This presentation is about predictability in climate change science. Towards the end of the 19th century, Svante Arrhenius recognized that carbon dioxide was going to accumulate in our atmosphere as a consequence of human activities, and furthermore, that it was going to cause a temperature increase at the surface of about 4 to 5°C, assuming that the concentration would double, a prediction that is remarkably close to what we think today. Earlier in the 19th century, the British scientist Tyndall already knew that the atmosphere would have a heat-trapping effect, and he carried out experiments to measure the infrared absorptivity of clean air. He was disappointed: clean air does not absorb infrared radiation. He later realized that carbon dioxide and water vapor do absorb efficiently this radiation, and that these gases are trace components of the atmosphere; as we now know, these are the most important greenhouse gases. However, the real pioneer was in fact Joseph Fourier, the French mathematician, who had concluded earlier that the amount of energy received from the Sun by the planets, and specifically the Earth as well, equals the amount of energy lost to space (Figure 1, see page 257). This energy emitted by the planets was called 'dark heat', *Chaleur Obscure* (infrared light had not yet been discovered). He did not know that the atmosphere would absorb this dark heat, and thus concluded that the surface of the Earth should be much colder than it actually is. We now know that without an atmosphere the temperature at the surface of the Earth would be about -18°C, and thus the oceans would be frozen. But the atmosphere plays a fundamental role as a warming blanket, and the average surface temperature is instead +15°C.

Carbon dioxide is the main greenhouse gas affected by human activities. It is biologically active; Figure 2 shows how it changes as a function of time and latitude, which is commonly called the pulse of the planet. In the winter its concentration increases due to respiration and in the summer

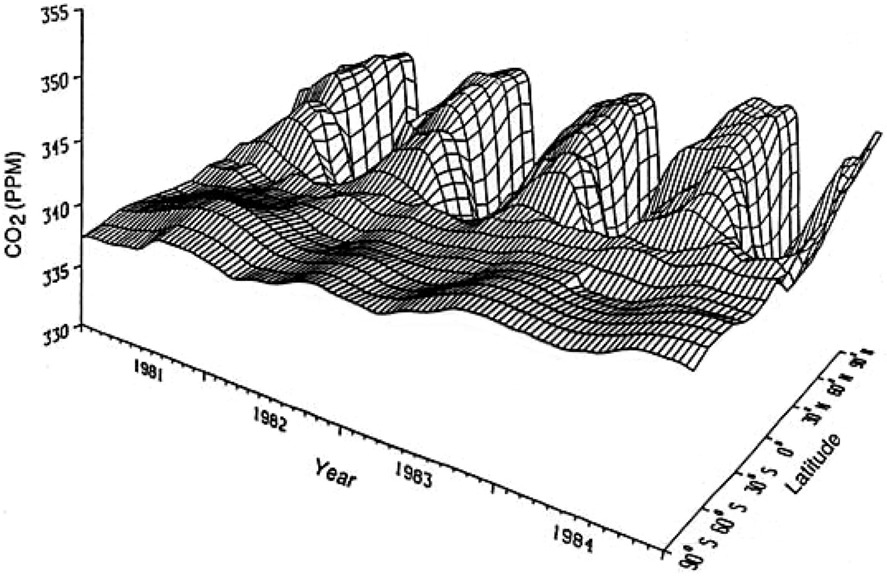


Figure 2.

photosynthesis is at work, absorbing carbon dioxide from the atmosphere. Thus, there is a balance that has been sustained for hundreds of thousands of years. Figure 2 also shows that the oscillations are more pronounced in the northern hemisphere, because of the topology of our planet: most of the continental mass is in the north. Figure 3 shows very clearly that the overall amount is increasing, because human activities are perturbing the balance of respiration and photosynthesis, as pointed out by Arrhenius.

Next, I will mention some results from a report by the Intergovernmental Panel on Climate Change (IPCC), an international group of scientists covering many different disciplines all having to do with climate. The last report (the third assessment report) was released in 2001, and a new one is going to be released in February 2007. There is no doubt that the chemical composition of the atmosphere is changing as a consequence of human activities, as shown in Figure 4 (page 79). Figure 5 (see page 258) shows the temperature change on a millennium time scale; this figure is the so-called hockey stick curve. The red lines represent an average of direct measurements of temperature, and the blue lines represent temperature inferred indirectly from other measurements such as the width of tree rings, coral reefs, etc., as

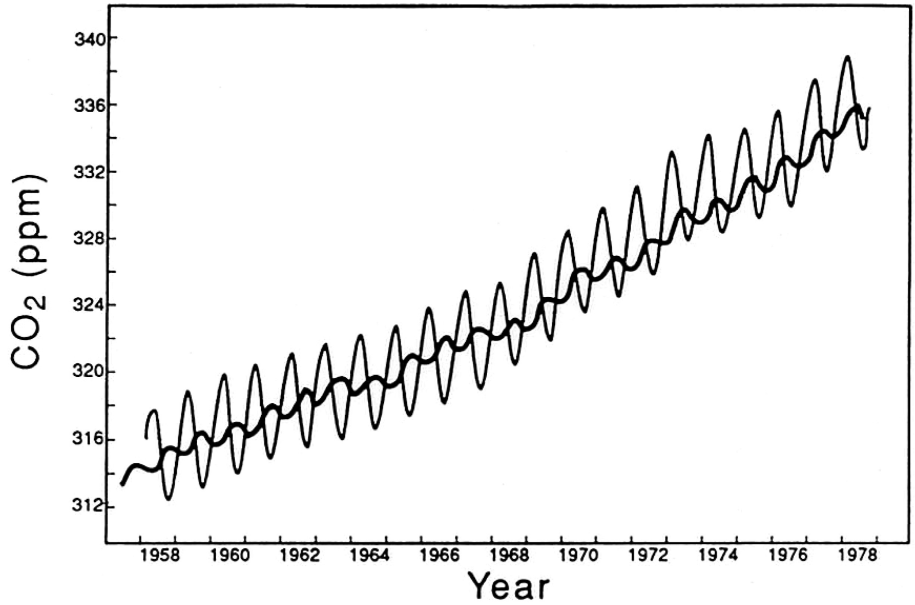


Figure 3.

there were no thermometers a thousand years ago. The important question addressed by the IPCC is the connection between temperature and composition changes: the curves have a similar shape. The conclusion is that there is indeed a connection. One way is to establish this connection is through atmospheric models: Figure 6 (see page 259) shows the results of a model simulation without taking into account the changes in the chemical composition as well as taking into account these changes. Accepting the connection makes it possible to make predictions of what will happen in this century, which depends on what society.

Figure 7 (see page 260) shows that 2005 has been the warmest year so far in the last 1000 years. The figure also shows that the temperature change in the tropics is smaller than in the poles. How has the climate changed on a geological time scale? We can go back half a million years, as shown in Figure 8 (see page 261); it is quite interesting that one can infer the composition of the atmosphere for such a long period of time by measuring the composition of air bubbles trapped in ice cores, and that one can also infer the temperature by measuring the isotopic oxygen concentration in the same ice cores. The oscillations in the figure indicate ice ages and

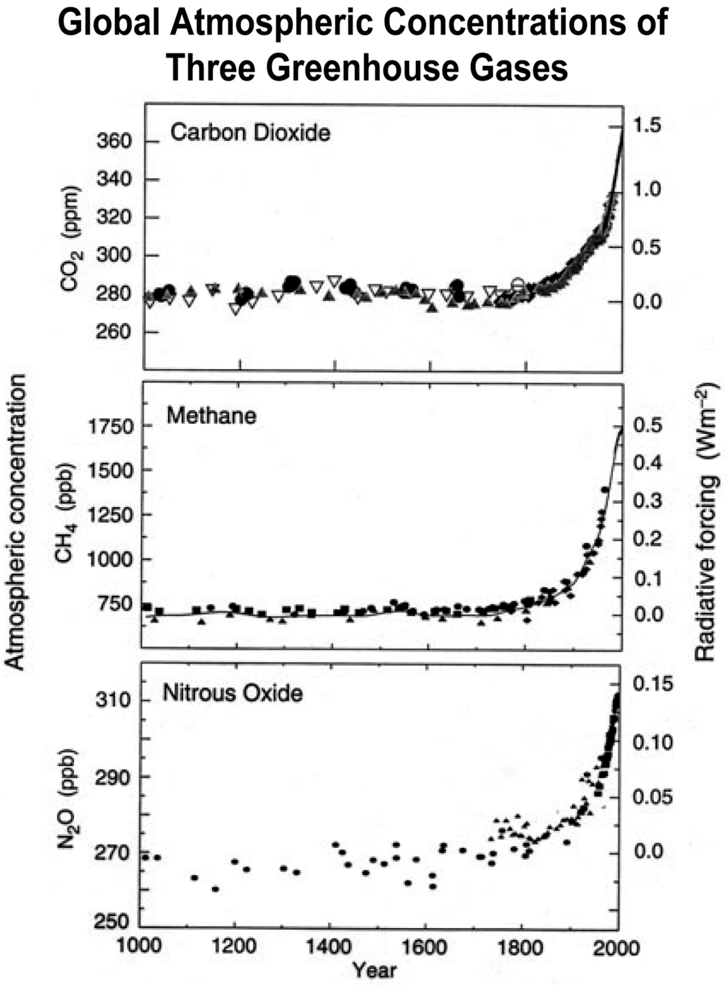


Figure 4.

interglacial ages. They are reasonably well understood: the explanation is given by the Milankovich theory, namely that the periodic temperature variations are a consequence of changes in the orbital parameters of the Earth, such as the tilt of the rotation axis and the eccentricity of the orbit, which change with periodicities of about 10,000, 40,000 and 100,000 years. Based on this understanding we can predict that we should have a reasonably stable climate period for the next thousand years. What we observe instead is a temperature spike taking place in just a few decades, well correlated with a similar spike in the concentration of carbon dioxide and other greenhouse gases. It is possible to carry out a statistical analysis and to estimate the probability that the recent global temperature change is a natural event versus an event connected with changes in the chemical composition of the atmosphere; the results indicate that there is about a 90% probability that the change is indeed related to human activities.

Observations indicate that the Earth's climate is clearly changing. For example, Glacier National Park (Figure 9, see page 261) will have to change its name relatively soon, as there will be no more glaciers in that Park. Yet another example is ice melting in Greenland (Figure 10, see page 262). If all of the ice over Greenland were to melt, sea level would rise about 6 meters. We have seen in recent years increased damage from intense hurricanes, such as Katrina (Figure 11, see page 262), or Wilma. Does this change have anything to do with climate change? Figure 12 (see page 263) shows the results of an analysis carried out by my colleague at MIT, Kerry Emanuel, indicating that the power of hurricanes is extremely sensitive to the surface temperature of the oceans, which has increased recently. It is not possible to establish that Katrina is necessarily a consequence of climate change, but one concludes, however, that statistically the strength of the hurricanes has increased as a consequence of human activities.

There is a now a consensus among scientists and climate experts that it would be very risky to have a temperature change of more than about 2°C or 2.5°C. To minimize this possibility it is necessary to stabilize the emissions of the greenhouse gases. The Stern Review released very recently involves mainly the economics of climate change, but it also summarizes the uncertainties. Depending on the levels of carbon dioxide that we end up having in the atmosphere, there is a 5 to 95% probability that the temperature will change by the amount indicated in Figure 13 (see page 263). So it is going to be quite difficult not to have a temperature change smaller than 2°C, which would require stabilizing carbon dioxide levels below about 550 ppm. If we do nothing, the 'business as usual' scenario, it is quite possible for the temperature to increase by more than 5°C, a change which begins to be comparable to the mean temperature difference between an ice age and an interglacial age.

In terms of climate change it turns out that it not only matters if carbon dioxide and methane atmospheric levels increase. Air pollution also matters, such as that induced by forest fires and transportation activities in cities. Figure 14 (see page 264) shows a satellite picture of fires in California, and Figure 16 shows savannah fires in West Africa; so we are burning large amounts of biomass in our planet. There is also a brown cloud, shown in Figure 15 (see page 264), described earlier by Professor Ramanathan. Figure 16 (see page 265), taken from the IPCC report, shows the relative contributions from different sources to climate change: first of all, the solar contribution is quantifiable and it is relatively small. The greenhouse gases, mainly carbon dioxide and methane, are the main contributors, but pollution of the type generated over large cities also contributes. The net effect of particles is actually quite complicated, but black carbon has a strong positive effect adding to global warming, as shown in Figure 17 (see page 265), taken from Jim Hansen.

Figure 18 (see page 266) compares the various sources of greenhouse gas emissions. Energy emissions are the largest, but there are also emissions from land use changes and agriculture. There is a way to reduce the energy-related emissions, as explained in Figure 19 (see page 266), taken from my colleagues Rob Socolow and Steve Pacala. The figure shows business as usual emissions going up relatively fast. In order to stabilize the atmospheric levels of greenhouse gases around 550 ppm the emissions have to be significantly reduced, and the figure shows schematically that a number of actions need to be implemented; no single action will do the job. The various actions that are needed are represented in the figure by 'wedges'. For example, more efficient buildings may be a little wedge; improved fuel economy and even nuclear energy are represented by other wedges, but cannot by themselves solve the problem. Another important wedge is carbon capture: instead of emitting to the atmosphere carbon dioxide from power plants, it can be pumped into oil wells or other cavities underneath the Earth's surface, so that it does not escape to the atmosphere. According to the Stern Review the cost of implementing these wedges is of the order of 1% of the global GDP, because the technologies required to address the problem in the next few decades are already available. However, it is very important to develop new technologies so that increasing amounts of energy are available in the future with reduced environmental impacts.

A major problem is that from the point of view of the developing nations, particularly India and China, it will be difficult to reduce emissions because their economies need to grow and this growth has traditionally been tied to an increase in energy consumption. For this reason it is essential for the developed countries to cut emissions to allow for some increase in the developing nations. However, even developing nations have to work

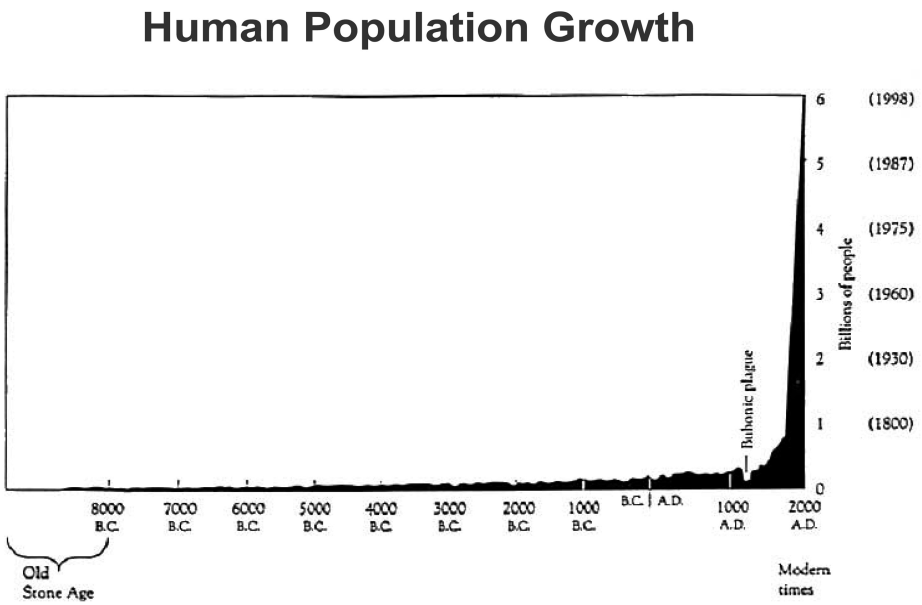


Figure 21.

very hard with energy conservation and with the implementation of the wedges mentioned above in order not to increase their emissions at the current rates (Figure 20, see page 267).

To finish, Figure 21 shows how the population of the planet has grown in the past few millennia. The current population is over 6 billion people. About one fourth of this population resides in the developed countries, which so far have been responsible for most of the pollution of the planet. The remaining three fourths have the right do develop their economies, striving to reach a standard of living comparable to that in the developed nations. The problem is, however, that the planet does not have sufficient natural resources and capacity to clean pollution from human activities to sustain such an economic growth if it takes place in the same manner as it took place in the past. Thus, it is imperative for developed and developing countries to work together to find new ways to achieve high standards of living in harmony with the environment.

ATMOSPHERIC CHEMISTRY AND CLIMATE IN THE ANTHROPOCENE*

PAUL CRUTZEN

I will speak about the Anthropocene, which is a new geological era created by mankind in the last 200 years or so, and I will also show you some results of the geoengineering exercises that we have been doing, basically to cool down the earth from too much heat caused by greenhouse gases.

Since the beginning of the 19th century we can agree that mankind has really opened a new geological era. Normally it is claimed we are in the Holocene, but we are no longer in the Holocene, we are in the Anthropocene because, in many ways, we determine the climate of the earth, its atmospheric chemistry and conditions at the surface. I shall give you some examples.

During the past three centuries human population increased by a factor of 10 and, in the last century which has just ended, by a factor of 4. Cattle population increased by 1400 million, that is, one cow per average family. Cattle produce methane, which we are interested in since methane is a greenhouse gas and also determines much of the background chemistry of the atmosphere. Urbanisation has grown more than tenfold in the past century and almost half of the people now live in cities, in megacities, and this tendency is increasing, especially in the developing countries. Industrial output increased 40 times during the past century and energy use 16 times, and almost 50% of the land surface has been transformed by human activity.

Water use increased by nine fold during the past century, to about 800 cubic metres per capita, most of which is used for irrigation, 25% for industry and 10% for households. To give some examples of the use of water resources, it takes about 20,000 litres of water to grow 1 Kg of coffee, 11,000 litres of water to make a quarter pounder and 5,000 litres of water to make 1 Kg of cheese: quite impressive, so no wonder we are running dry.

* This is a transcript of the author's lecture during the Plenary Session, which the PAS is publishing as is, without the author's corrections.

Another example of human activities is the appropriation of terrestrial net primary productivity: it looks like mankind is using about 30% of the natural resources available in terrestrial net primary productivity. Fish cash increased 40 times, the release of sulfur dioxide to the atmosphere, which only 2 decades ago was 160 Tg/year (a teragram is 10^{12} grams or 1 million ton), is fortunately now down to 110 Tg/year. There has been an improvement, and that is because sulfur emissions had caused major problems, for instance acid rain, bad effects on health, poor visibility, and also have an impact on cloud formation and sulfate aerosol formation. Release of NO to the atmosphere from fossil fuel and biomass burning is larger than its natural inputs, causing high surface ozone levels over extensive regions of the globe.

However, several climatically important 'greenhouse gases' have substantially increased in the atmosphere: carbon dioxide by more than 30% and methane by more than 100%. Most of these changes have actually taken place or picked up since the end of the last world war, so this is what we call the *Great Acceleration* (Fig. 1): for instance, population increase, total real gross domestic product, foreign direct trade, the damming of rivers, which is a major activity of mankind and, also, the growth of McDonald's around the world which, of course, has to do with methane release and the involvement of cows. I could mention many more examples but these may suffice.

Humanity is also responsible for the presence of many toxic substances in the environment, even some that are not toxic at all but that have, nevertheless, led to the ozone hole, and those are, of course, the chlorofluorocarbon gases. CFC gases are very inert in the troposphere, are destroyed by ultraviolet radiation above about 25 km in the atmosphere and then give rise to chlorine atoms that break down in the stratosphere. This was proposed and hypothesised for the first time by Mario Molina, who is in the audience. They also cause UV-B radiation and skin cancer.

We also have species extinction. The natural extinction rate of species was roughly 1 species per million species per year: it is now about a factor a thousand time larger, so the average life span of species in the atmosphere is close to one thousand times shorter than in pre-industrial, pre-Anthropocene conditions.

Regarding erosion, we are experiencing 15 times the natural erosion rate as a result of human activities, man-caused erosion, crop tillage, land conversion for grazing, and construction. So, at the current rate, anthropogenic soil erosion would fill the Grand Canyon in about 50 years. We are disturbing the nitrogen cycle: here you can see (Fig. 2, see page 86), as a function of time, the natural nitrogen fixation rate by leguminous plants,

The great acceleration

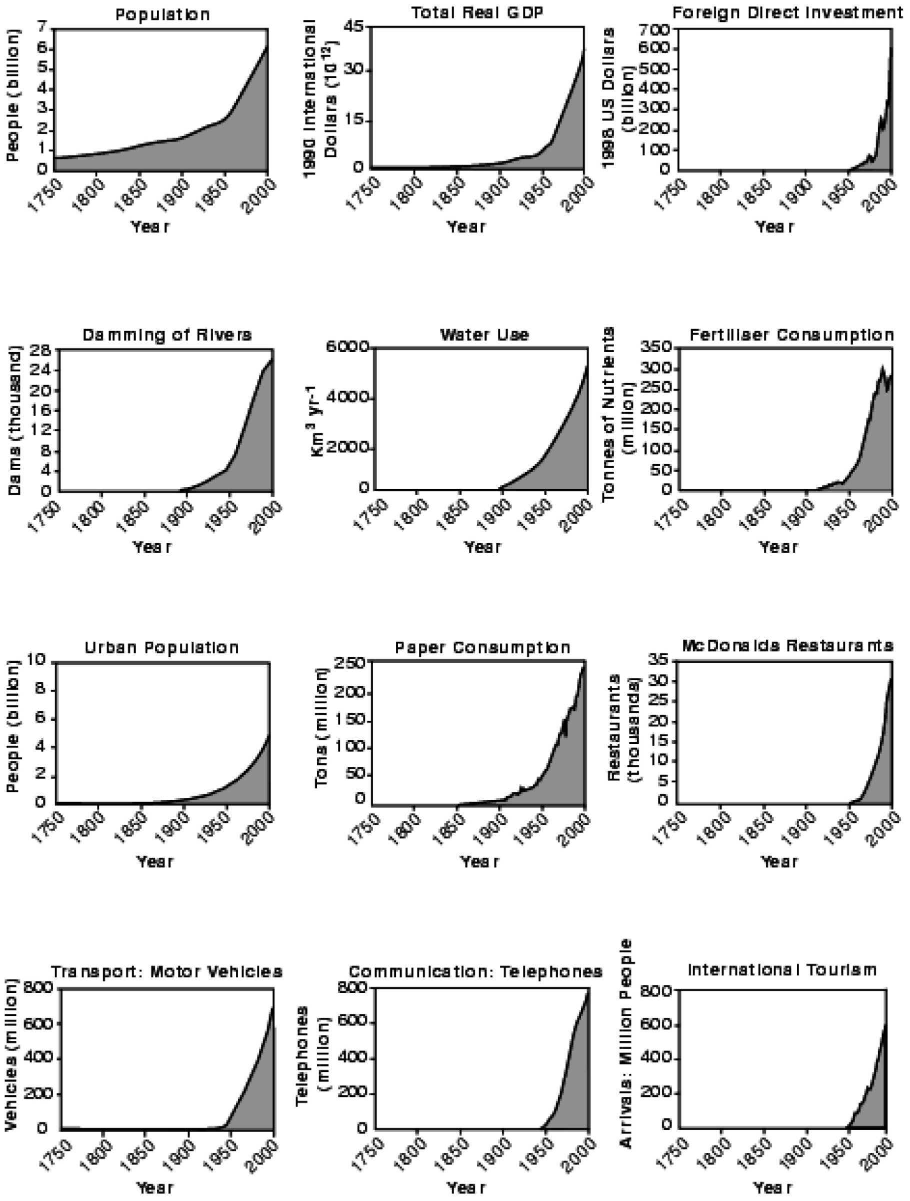


Figure 1.

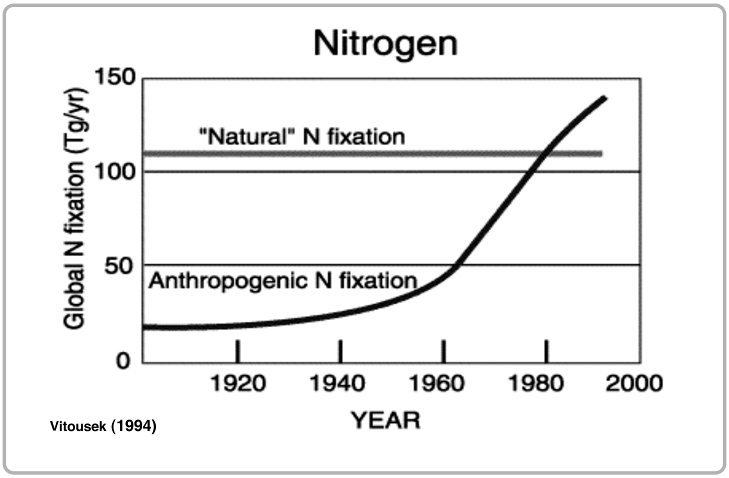


Figure 2.

and also by lightning, and you can also see the input by anthropogenic activities. Since about 1980 the anthropogenic nitrogen fixation, mainly fertiliser nitrogen, has been bypassing the natural N fixation rate, and that has major consequences for the emissions of nitrous oxide in the atmosphere. I will come back to that. It is amazing that anthropogenic nitrogen fixation is growing, although when we really look at what ends up in the mouths of people, that is only in the order of 10%, so 90% is sort of wasted. It really is a pity that agriculture is so inefficient in using its nitrogen.

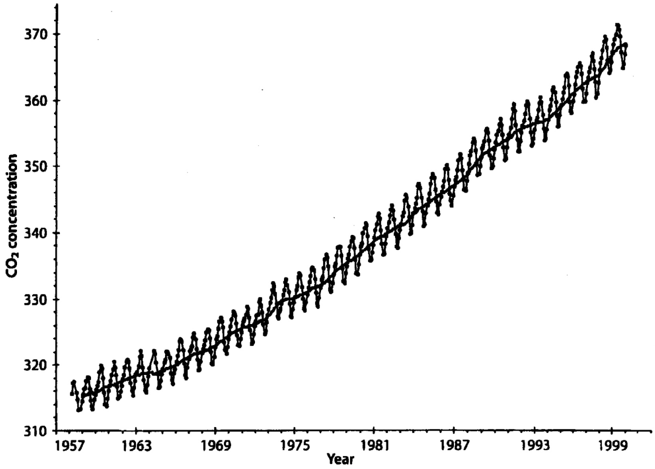
The composition of the atmosphere is dominated, as you all know, by nitrogen, oxygen, argon; we do not have to worry that we will run out of those, we leave the study of those compounds mainly to geologists. Atmospheric chemists and climate researchers nowadays are much more interested in the minor constituents in the atmosphere, starting with carbon dioxide, of which we now have 380 ppm (the numbers in the figure are a little outdated). It is growing by about 0.4% per year, it is, of course, the major greenhouse gas and, of course, it is involved in the biosphere and photosynthesis. It does not play a major role in the chemistry of the atmosphere. There we go down to gases with even lower concentrations, like methane, of which we have about 1.7 ppm in the atmosphere. This is double the amount of pre-industrial times, it has been growing quite considerably but, at the moment, is at a standstill, and I will briefly come back to that. Ozone is an extremely important trace gas, at ground level it has both positive and negative effects. The positive effect at ground level is that it promotes the production of hydroxil radicals that clean the atmosphere. The bad effect is that, if there is too much ozone in the surface air that we breathe, it is not very healthy for people. In the stratosphere we have seen a decline of ozone because of the use of chlorofluorocarbons, which I will briefly discuss. It is very variable: we have, on average, about 30 parts per billion of ozone in the troposphere, while in the stratosphere we can have in the order of 10 millionth of ozone.

Nitrous oxide is a by-product of the nitrogen cycle. We now have about 0.32 ppm of nitrous oxide in the atmosphere and it is growing by about 0.25% per year. Then we have CFC gases. They are no longer growing, they are actually going down now very very slowly in the atmosphere because there has been international agreement to stop their production so the ozone layer will slowly recover.

I will show you some viewgraphs of the major effects of human activities (Fig. 3, see page 88): in the upper figure you see a steady rise of the carbon dioxide amount in the atmosphere. These data go back to the end of the 1950s, by Dave Keeling, who unfortunately died last year but left this record behind. You can also see the seasonal variations in carbon dioxide, having to do with photosynthesis, for instance. Then you see the ozone hole pictures below and you can see, in the picture on the left, a very drastic depletion in total ozone, since the early 1970s, over Antarctica, especially in the springtime month of October. If you look at the right hand side of this picture you see that the ozone destruction (what is shown there is the ozone concentration as a function of altitude) is especially happening at an altitude where we normally have a maximum of ozone. Within two months that ozone maximum had collapsed into an ozone minimum.

This was something that had never been predicted, so we have an example of what Prof. Zichichi was talking about, the data had to be collected, and initially the observers did not believe that data because this was so unexpected, something like this happening at the other side of the world where the CFC gases were not at all injected into the atmosphere. It is basically radical reactions, chlorine atoms which enter into catalytic reactions destroying ozone. Hence, for each chlorine atom produced from the CFCs, you destroy up to 100,000 ozone molecules before the chlorine is removed from the atmosphere. This is quite shocking, mankind has created a chemical instability in the atmosphere by the use of CFC gases, which look so innocent because you can even breathe them at ground level and they will not harm you very much, but this is what they are doing in the stratosphere.

The "Keeling curve", showing the steady increase in atmospheric CO_2 concentration recorded monthly at Mauna Loa in Hawaii, 1958–1999 (adapted from Keeling and Whorf 2000)



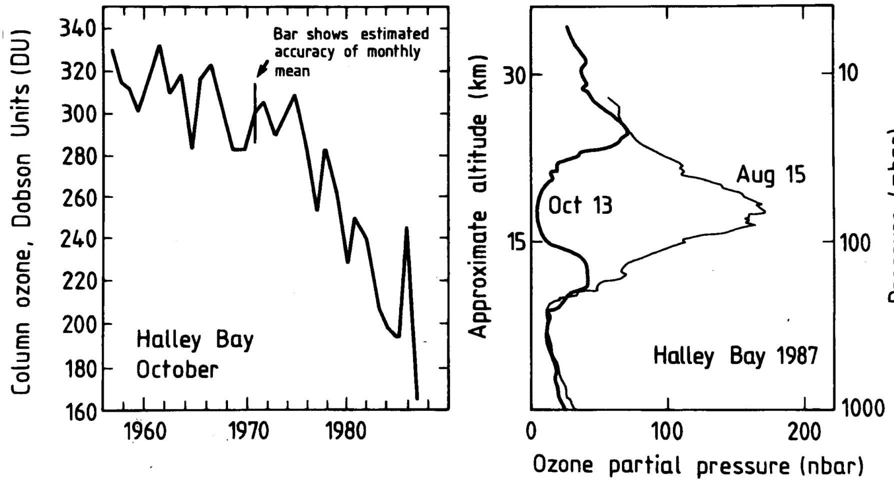


Figure 3.

Here we have figures showing the greenhouse effect (Fig. 4, see page 89). I will be very brief here. The earth is supplied by the sun, on average, by about 340 W/m^2, let us call that 100 units: of these 100 units almost 30 are scattered back to space, that is, by reflection at cloud tops and also by scattering of particles in the atmosphere. I will come back to that. There is also some absorption in the atmosphere in clouds and also elsewhere in the stratosphere by

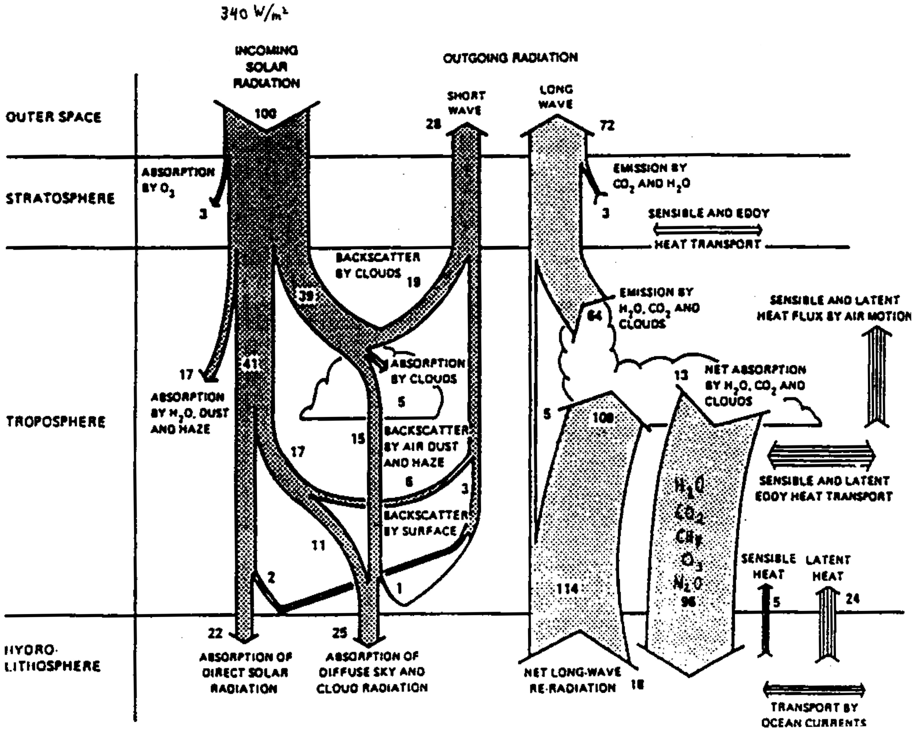

340 W/m²
INCOMING SOLAR RADIATION
100
OUTGOING RADIATION
SHORT WAVE
28
LONG WAVE
72
OUTER SPACE
STRATOSPHERE
TROPOSPHERE
HYDRO-LITHOSPHERE
ABSORPTION BY O_3
3
EMISSION BY CO_2 AND H_2O
3
SENSIBLE AND EDDY HEAT TRANSPORT
BACKSCATTER BY CLOUDS
19
39
EMISSION BY H_2O, CO_2 AND CLOUDS
64
SENSIBLE AND LATENT HEAT FLUX BY AIR MOTION
17
ABSORPTION BY H_2O, DUST AND HAZE
41
ABSORPTION BY CLOUDS
5
NET ABSORPTION BY H_2O, CO_2 AND CLOUDS
13
5
100
17
15
BACKSCATTER BY AIR DUST AND HAZE
6
3
SENSIBLE AND LATENT EDDY HEAT TRANSPORT
H_2O
CO_2
CH_4
O_3
N_2O
96
BACKSCATTER BY SURFACE
11
2
1
114
SENSIBLE HEAT
5
LATENT HEAT
24
22
ABSORPTION OF DIRECT SOLAR RADIATION
25
ABSORPTION OF DIFFUSE SKY AND CLOUD RADIATION
NET LONG-WAVE RE-RADIATION
18
TRANSPORT BY OCEAN CURRENTS

Figure 4.

water vapour, so what comes down to the earth's surface is about 47 units, the sum of 22 and 25. The earth has to get rid of this energy because, otherwise, in the space of a few weeks, it would start boiling around us. How is it done? It happens by the release of latent heat at the earth's surface, 24 units; 3 units sensible heat release, which added to 24 makes 27 units, and 47 units that we have to get rid of, so there are 18 units left. But now, in a very miraculous way, the earth and its atmosphere is taking care of the release of these 18 units, because it is not just emitting 18 units in the atmosphere, but it is emitting 140 units of which about 96 units are coming back, and that is due to the fact that the earth has the greenhouse gases in the first place, water vapour and carbon dioxide, but also methane, nitrous oxide, and ozone. The CFC gases are all contributing to this back flux of infrared radiation. So of the energy supplied by the sun, of the heat, there is about a factor of 6 recycling of energy taking place, which, of course, makes life on earth possible at all, and which is due to the greenhouse gases, many of which have anthropogenic sources like CO_2 as I just mentioned.

The temperatures on earth really are rising (Fig. 5, see p. 268). You can see a steady rise in temperatures since 1970. The rise in temperature that we already had earlier may have had to do with natural variability of climate and maybe with some solar activity, but there is no doubt that nowadays, since 1970, we have had a steady rise in the temperatures in the atmosphere, as a global average, and it is continuing. Five or six of the warmest years were in the last decades, a clear sign that something is happening.

But the greenhouse gases are not the only factor that we have to consider. They heat the earth but there are also factors which lead to its cooling, and those are particles in the atmosphere (Fig. 6, see p. 269).

Now the uncertainties. The particles, many of which are released by human activities, cool the earth and also serve as cloud condensation nuclei, so they make the clouds more reflective to solar radiation, which has a cooling effect. The uncertainties, however, are very large, so we have to improve on that, but it will be a very difficult process to really estimate accurately what is the contribution of aerosol particles in the atmosphere. Many of these aerosol particles are produced by human activities, air pollution, and have a damaging function when you breathe them, so we really enter here into a dilemma because we want to get rid of these particles, due to their effects on health, but, by doing that, we increase the heating of climates, because the reflection of solar radiation to space is diminishing. This is a dilemma for policy-makers and, of course, us scientists and the general public. We have tried to estimate a little bit what the energy balance may be: the average amount of heat supplied by the sun to earth is 340 W/m^2, the greenhouse forcing is 2.7 W/m^2, the heating of the ocean is subtracting about 0.3 W/m^2, and, also because temperatures on earth have increased by about 0.6 to 0.7° in the atmosphere, we have an increased release of energy to space in the infrared by 1 W/m^2. But these factors combined, the 1.3 W/m^2, do not balance the 2.7 W/m^2 of heating, which means there is 1.4 W/m^2 left that the earth has to get rid of. Prof. Ramanathan and I, who did this analysis, think this is due to increased albedo effect, higher reflectivity of solar radiation by the clouds and, also, in general, the reflection of sunlight on the particles. Now it is interesting, if you improve the conditions, if you remove the particles from the atmosphere, meaning that the 1.4 W/m^2 becomes 0, basically, the net heating of the atmosphere will increase much more, about double as much compared to what is estimated here. It is instructive to note that the pure release of heat to the atmosphere by the burning of fossil fuels is only 0.025 W/m^2, that is only 1% of the greenhouse forcing, so the real problem is the greenhouse gases, it is not the heat we put in the atmosphere, that has a minor effect. In fact, the heat released from earth is larger, 0.087 W/m^2, than the heat released by the burning of fossil fuels.

So what are the effects? It is clear. The Intergovernmental Panel for Climate Change, under the auspices of the United Nations, brings out, every six years, its estimate of the situation, and what it said in 2001 (another report is due next year), is that there is new and stronger evidence that most of the warming observed over the last 50 years is attributable to human activities. The rise in temperatures during this coming century may be between 1.4 and 5.8°C, a very large uncertainty that has to do with uncertainty in science but also in the human behaviour in the future. How much fossil fuel are we going to burn in the future? All this leads to a major uncertainty.

These are big numbers, even the lower number here is quite substantial and will have effects on the earth's climate. It will cause the sea level to rise, estimated to maybe up to 80 cm, almost a centimetre per year in this century, since there are some signs that the upper limit of this range may be closer to the truth than the lower limit, there will be redistribution of precipitation, and, for instance, Italy and the Mediterranean regions will get substantially drier, which is something to worry about. Northern Africa I do not have to mention, lots of people are already moving away from Africa to Europe, enhanced risk of extreme weather, flooding, desertification, we had this very hot summer in 2003 and, again, this year is very odd from the point of view of meteorology. Too-rapid changes in temperatures will cause that the ecosystems cannot adapt to the situation.

What should we do about it? Well, in the first place, we should reduce emissions of greenhouse gases in the atmosphere, but that is easier said than done, because, to stabilise the amount of CO_2 in the atmosphere, we would have to reduce emissions by more than 60%, and that is not taking into account the growing contribution by developing countries, so will we ever be able to achieve the 60% or more reduction? One can be rather pessimistic, unfortunately. Methane I already mentioned: at the moment we do not see any increase in methane in the atmosphere for a while. That does not mean that it will not come back in the future, because with higher temperatures the permafrost regions in the northern latitudes, Canada and Russia, will thaw, which may lead to emissions of methane and carbon dioxide in the atmosphere and increased warming. A 70 to 80% reduction in the emissions of nitrous oxide would be required to stabilise its amount in the atmosphere, and this has to do with food production, nitrogen fertiliser, and I do not see that happening at all.

Fortunately, we have some success stories: CFC gases are no longer produced, only in very small amounts. But, nevertheless, you may have heard that this year was a very bad year for the ozone hole, the deepest ozone hole was just this year, despite the fact that these gases are very slowly disappearing from the atmosphere. But the activity of the CFC gases is also very much dependent on temperatures in the atmosphere. Clouding can only be activated if you have ice particles in the atmosphere, if you have this at higher latitudes, and this was a very cold year in the Antarctic, therefore causing very large ozone depletions. I am sure Mario Molina will go into detail.

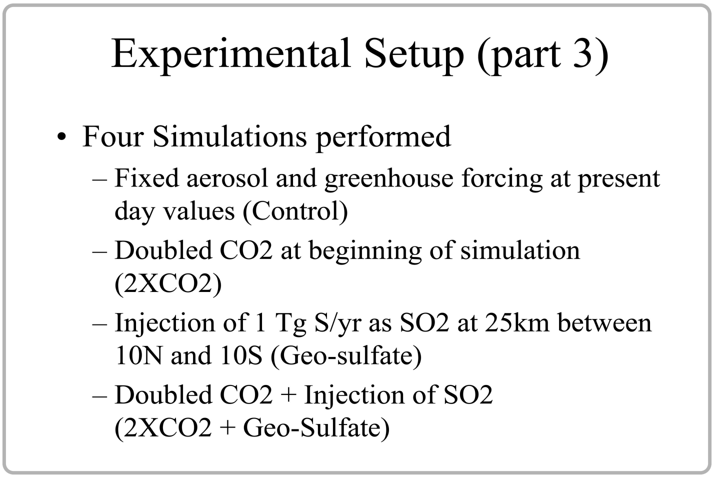
In the reduction of the greenhouse gases, here we can see (Fig. 7, see p. 269) emissions in metric tons per capita per year, leading in North America, Oceania, Europe and you see all the developing countries here, which are hardly part of the game but they will be part of the game because they want to increase their standard of living. We should not believe that nature will help us out, that when temperatures go up then the greenhouse gases will go down cooling the earth, no, it is the opposite, when temperatures go up CO_2 also goes up at the same time, and methane goes up causing climate variability, because the Milankovich Cycle is supported or enhanced by the natural emission of greenhouse gases. Most sensitive are the higher latitude regions, where, for a doubling of the amount of CO_2 in the atmosphere, we can have temperature increases in the order of 6 to 8°C, and that may lead to the thawing of the permafrost regions which I have already mentioned, which would create a positive feedback effect enhancing temperatures even more than just by temperature changes.

New studies indicate that the Arctic Ocean ice cover is about 40% thinner than 20 to 40 years ago, and there is dramatic climate change happening in the Arctic, about 2 to 3 times the pace for the whole globe and so this may lead to what I have already mentioned, to melting of the permafrost and another major positive feedback factor (Fig. 8, see p. 270).

Can we do something about it? If you say, reduce the emissions of greenhouse gases, and the way to go is get your energy from other sources, in the first place by energy savings, there is a lot that can be done there, renewable energy, nuclear energy, wind and solar power, and CO_2 sequestration is also a possibility. There is another possibility and that is to inject sulfur in the stratosphere, by bringing, for instance, H_2S with rockets and balloons in the stratosphere where you oxidise H_2S to SO_2, which is further oxidised to sulfuric acid, which forms sulfate particles that reflect sunlight. So this is, in principle, possible to do and I will show you the results of some calculations. I do this work in collaboration with scientists from the National Centre for Atmospheric Research, Philip Rasch and D.B. Coleman. The concept of doing this geoengineering goes back to Budyko and the study in the National Academy of Sciences in 1992, and more recent studies, the Teller proposal, then Govindaswamy and Caldera, and then I did a study which came out in August of this year. We use a General Circulation Model to estimate what might happen if you put some sulfur into the stratosphere in the form of sulfate particles that reflect solar radiation. We use a model with rather complete physics and simple chemistry but no biological feedback, so the permafrost story is not included in the model, also we hardly have a Carbon and Nitrogen Cycle. The circulation of the ocean is not changed because of our emissions of sulfur gases in the atmosphere.

Some further information about the model. The model goes up to 80 km altitude in the atmosphere, it is 52 layers, and has a special distribution of 2.5° by 2° latitude and longitude and about less than 1 km altitude speciation.

So now we do some experiments. We basically conduct four simulations (Fig. 9). In one simulation we work with the current atmosphere, fixed aerosol and greenhouse forcing, as happens at the moment in the atmosphere. Then, with the same initial conditions, we double the amount of CO_2 and then look at what happens to the average tempera-



Experimental Setup (part 3)

- Four Simulations performed
  - – Fixed aerosol and greenhouse forcing at present day values (Control)
  - – Doubled CO2 at beginning of simulation (2XCO2)
  - – Injection of 1 Tg S/yr as SO2 at 25km between 10N and 10S (Geo-sulfate)
  - – Doubled CO2 + Injection of SO2 (2XCO2 + Geo-Sulfate)

Figure 9.

ture; we already know the answer: it will go up. Then we do an experiment in which we inject 1 million Tg S/yr as SO_2 in the atmosphere at near 25 km altitude, in between 10°N and 10°S and that should lead to cooling, and I will show you that it does. And then we do both, we double the amount of CO_2 and we inject SO_2 and we look at the net result of that. Here (Fig. 10, see p. 271) we have the net result. The basic case, the control case is basically here, you can see it in yellow, but the model would predict average global temperatures of the order of a little above 288°K. If you double the amount of CO_2, but you do nothing with the sulfate particles, you see the rise in temperatures, a little over 2°C is the rise in temperatures starting from these conditions. If you put the sulfate particles in the stratosphere you go here, you have the cooling, also by about 2°C or more. If you do both, you end up with the black curve, you are almost back to the unperturbed conditions.

What I showed you before with temperatures, you can also see with precipitation (Fig. 11, see p. 271): doubling of CO_2 gives more rainfall, more particles in the atmosphere, less rainfall, all in mm/day and then you do both and you get basically the same initial state back. So, if the amount of CO_2 in the atmosphere doubles and temperatures go up, then we have the possibility, by adding sulfur to the stratosphere, to come back to normal temperature conditions.

The lifetime of the aerosol particles in the stratosphere is of the order of three to four years (Fig. 12, see p. 272). Normally it is shown that it is 1 to 2 years but, when you do the model calculation, it comes out to much longer. That means emissions into the stratosphere do not have to be as high as would otherwise be the case. The optical depth of the sulfate is about 0.06, which means that the sky will become a little lighter, but, on the other hand, you will also get wonderful sunsets and sunrises. It is basically a human volcano which is produced here. Precipitation changes around the globe are not very large, the average for the globe is 2.8, so the maximum deviation in this case would be 0.5, but these data are statistically not significant. On the whole, you can say that the precipitation changes have been less than 10% of the normal precipitation.

We can look at the temperatures (Fig. 13, see p. 273). If you do the doubling experiment of CO_2 you see the heating of the higher latitude regions of the atmosphere by the greenhouse effect, if you do doubling of CO_2 at the same time as sulfate in the atmosphere you get the white colour almost everywhere, meaning temperature changes between -1 and 1°C all over the globe, of the order of a few tenths of temperature changes in the atmosphere.

Let me stop here, we should definitively leave some time for discussion. This is, of course, an experiment which, at the moment, you only do on a computer. If we ever have to do a thing like that, it will only be if our climate runs away in some way. So you might think that it is very unlikely, but we have seen, in the case of the ozone hole story, that very unlikely things might happen, so we should be prepared by surprises in the future, which can only be discovered again by observations. The models are getting better. There is already considerable theoretical work going on concerning this sulfate experiment, more people are coming into action; until a few years ago or, basically, until this year this was a taboo thing, you should not study a thing like that, but this has changed now. There is a lot of activity taking place, you only do this when things are really getting very bad but we better be prepared that, if that is the case, in order to have some kind of weapon available to reduce the bad effects of other human activities.

Here I would like to stop: I am quite certain there will be very critical remarks but I thank you for your attention.

SESSION III

PREDICTIONS IN THE LIFE SCIENCES

PREDICTING THE MINIMAL SUSTAINABLE GENOME

RAFAEL VICUÑA

Predictability is the capacity to foretell a given situation or event based on observation and reasoning. This ability is fundamental to the progress of science, since it relates to the processes of deduction and induction. Predictability is in itself a test of scientific understanding, since the power of scientific method shows its strength when predictions made by investigators coincide with observable events in the laboratory or in nature.

Recent advances in functional genomics and computational biology have led to considerable progress in our comprehension of gene structure and expression. One of the most defying issues that geneticists are presently addressing is the so-called minimal genome concept, in other words, the attempt to define the minimum number of genes that are necessary to sustain a free-living cellular organism. It is expected that the lowest number of genes will be identified under the most favorable conditions imaginable, that is, in the presence of a full complement of nutrients and in the absence of environmental stress. Until the size of such minimal genome is experimentally confirmed, this task epitomizes an exercise of predictability.

The ability to correctly define the minimal gene set goes to the heart of our understanding of cellular life. In addition, the fulfillment of this goal should provide some insight into the earliest stages of biological evolution, since it is assumed that simpler free living cells, with genomes much smaller than those of extant microbes, must have proliferated at the onset of life on earth. On the other hand, if a minimal genome proves to be something that we can observe in nature or obtain in the laboratory, some may be tempted to believe that the challenge of synthesizing simple forms of cells may be plausible in a not so distant future.

The Minimal Cell

Closely related to the minimal genome problem is that of the minimal living cell, defined as a theoretical entity having the smallest number of components and functions that are necessary to be considered alive.[1,2,3] Although it might not be straightforward to reach an agreement on the meaning of 'alive' in this context, this minimal cell would need to exhibit the properties of self-maintenance or metabolic activity, self-reproduction and Darwinian evolution.[3] A pioneering effort to describe a minimal cell was achieved by Morowitz.[4] Based on enzymatic reactions that he considered essential, Morowitz arrived at the conclusion that such minimal cell should be about one tenth the size of Mycoplasma, a very small and well described bacterial pathogen. However, it is most likely that primitive cells were even simpler than the one predicted by Morowitz and that complexity gradually built up as a result of billions of years of evolution of new metabolic pathways and defense mechanisms.

In a recent publication, Szostak *et al.*[5] have envisaged how an early protocell might have looked like. The nucleic acid could have been a double stranded RNA molecule, with one strand possessing replicase activity and the complementary strand serving as its template. Physical confinement of this genome was necessary to facilitate preferential replication of those RNA molecules incorporating advantageous mutations, thus allowing Darwinian evolution. Self-assembling and self-replicating vesicles composed of amphipathic lipids must have been responsible for encapsulating the RNA molecules. At this point, coupling between membrane synthesis and replication of the genome would be required to improve survival of the entire entity. This condition could have been met by a second gene encoding a ribozyme catalyzing the synthesis of phospholipids necessary to build up the membrane.

Another proposal of a minimal cell, in a closer agreement with contemporary life, came as a result of the so-called E-CELL project.[6] The proposed

[1] Luisi, P.L., Oberholtzer, T. and Lazcano, A., 'The notion of a minimal cell: A general discourse and some guidelines for an experimental approach', *Helv. Chim*. Acta 85, 1759-1777, 2002.

[2] Islas, S., Becerra, A., Luisi, P.L. and Lazcano, A., 'Comparative genomics and the gene complement of a minimal cell', *Origins of Life and Evolution in the Biosphere* 34, 243-256, 2004

[3] Luisi, P.L., Ferri, F. and Stano, P., 'Approaches to semi-synthetic minimal cells: a review', *Naturwissenschaften* 93, 1-13, 2006.

[4] Morowitz, H.J., 'Biological self-replicating systems', *Progr. Theor. Biol*. 1, 35-58, 1967.

[5] Szostak, J.W., Bartel, D.P. and Luisi, P.L., 'Synthesizing life', *Nature* 409, 387-390, 2001.

[6] Tomita, M., K. Hashimoto, K., Takahashi, K, Shimizu, T.S., Matsuzaki, Y., Miyoshi, F. Saito, K., Tanida, S., Yugi, K., Venter, J.C. and Hutchison III, C.A., 'E-CELL: software environment for whole-cell simulation', *Bioinformatics* 15, 72-84, 1999.

autonomous cell in this case required a total of 105 protein-coding genes. These were involved in very rudimentary metabolic pathways for glucose metabolism and phospholipid biosynthesis, in gene transcription and in protein synthesis. Accordingly, this minimal cell should be able to maintain metabolic homeostasis, but could not reproduce or evolve since it could not synthesize DNA with the proposed genome.

Theoretical Approaches to the Minimal Genome

However, although highly provocative, this hypothetical minimal cell is pure speculation. The simplest living cells today possess a few hundred DNA genes encoding a corresponding number of proteins. Most of these proteins are catalysts of numerous reactions taking place in a highly confined and organized environment. There are also some genes in the cell that encode various types of RNAs. Table 1 illustrates some small genomes found in bacteria, most of which are endosymbionts and obligate parasites. For comparative purposes, the genome sizes of *Escherichia coli* and *Streptomyces coelicolor* are also shown.

The hyperthermophile *Nanoarchaeum equitans* possesses the smallest genome that has been sequenced and analyzed.[7] It is an obligate symbiont that grows in co-culture with the crenarchaeon *Ignicoccus* and therefore, as opposed to *Mycoplasma* and *Haemophilus*, it cannot grow in the laboratory as a free-living microorganism. In contrast to other microbial genomes that are undergoing reductive evolution, *N. equitans* has an unusually high gene density, with little non-coding DNA and very few pseudogenes. In addition to its 556 protein coding genes, it has 17 genes encoding ribosomal RNA. Adaptation to an obligatory parasitic life is evidenced by the lack of most genes involved in *de novo* synthesis of aminoacids, nucleotides, cofactors and lipids, as well as genes of the glycolytic/gluconeogenesis pathways, the pentose shunt and the Krebs cycle. However, in contrast to other parasites, it contains most of the enzymatic machinery required for DNA repair and the complete machinery for DNA replication, transcription and translation. The lack of genes encoding transfer RNAs and the large amount of coding capacity devoted to surface proteins that interact with the host seem to leave little room for further reductions of the genome.

[7] Das, S., Paul, S., Bag, S.K. and Dutta, C., 'Analysis of *Nanoarchaeum equitans* genome and proteome composition: indications for hyperthermophilic and parasitic adaptation', *BMC genomics* 7, 186, 2006 (Epub).

TABLE 1. GENOME SIZE OF SOME PROKARYOTES

Microorganism	Genome size (kb)	Protein coding genes
Mycoplasma genitalium	580	482
Nanoarchaeum equitans	491	556
Buchnera aphidicola BBp	616	545
Blochmannia floridanus	706	625
Chlamydia tracomatis	1,000	895
Rickettsia prowazekii	1,100	834
Aquifex aeolicus	1,591	1,553
Haemophilus influenzae	1,830	1,703
Escherichia coli	4,640	4,288
Streptomyces coelicolor	8,500	7,825

Some other very small genomes found in nature are those of bacteria living in endosymbiosis with insects. These include *Blochmannia floridanus*, *Wigglesworthia glossinidia* and *Buchnera aphidicola*, the endosymbiotic bacteria of carpenter ants, tsetse flies and aphids, respectively. In these obligate, metabolic interdependent insect-bacterial relationships, the bacterial cells are contained in specialized host cells called bacteriocytes and the infection is vertically transmitted via eggs and young embryos. These endosymbionts diverged from their free-living relatives approximately 70 (*Blochmannia*) to 250 (*Buchnera*) million years ago. Since then, they have experienced a genome minimization process that started with a rapid decline and gradually slowed down until reaching the sizes found in present day genomes, which are 20-25% of the original ones. There are species of the genus *Buchnera* possessing genomes as small as 450 kb, much smaller than that of *Mycoplasma genitalium*.[8] However, they have not been sequenced and annotated to date. On the other hand, the hyperthermophile *Aquifex aeolicus* possesses the smallest sequenced genome of an autotrophic microorganism.

[8] Gil, R., Sabater-Muñoz, B., Latorre, A., Silva, F.J. and Moya, A., 'Extreme genome reduction in *Buchnera* spp.: Toward the minimal genome needed for symbiotic life', *Proc. Natl. Acad. Sci. USA* 99, 4454-4458, 2002.

The question is whether the very tiny genomes shown in Table 1 can be further reduced without affecting life's viability. Several laboratories have approached this problem, both theoretically and experimentally.

Some years ago, Mushegian and Koonin made a comparison between the genomes of two parasitic bacteria, namely *Mycoplasma genitalium,* a Gram positive human urogenital pathogen with 482 open reading frames (ORFs), and *Haemophilus influenzae,* a Gram negative human parasitic bacterium causing pneumonia and other diseases, with 1,700 ORFs.[9] The rationale followed by these investigators was that the genes that are conserved in these two bacteria are almost certainly essential for modern-type cellular function and therefore likely to approximate the minimal gene set. The authors found 240 *M. genitalium* genes have orthologues[10] in the genome of *H. influenzae*, although at first sight this resulting assemblage appeared insufficient since some key enzymes from intermediary metabolism were missing. The reason for this presumed incompleteness, which has also been observed by other authors in similar studies, is the so-called nonorthologous gene displacement (NOGD). This implies that unrelated or distantly related proteins are adapted to perform the same function in their respective cells. After identifying 22 nonorthologous displacements, the authors selected them from the genome of *M. genitalium* and they were added to the set of shared orthologues. Of the resulting 262 genes, 6 were eliminated from the minimal set because they were likely to be specific for parasitic bacteria. The conclusion was then reached that a hypothetical microorganism with 256 genes would be minimally equipped with the required cellular components, metabolic pathways, systems to copy and express its genome, a signal transduction apparatus and a few chaperones. It must be kept in mind, however, that genomes also contain some sequences encoding RNA species such as ribosomal RNA or transfer RNA, which are definitely essential for life. Therefore, these non-protein coding genes should be included in a minimal gene set. In the case of *M. genitalium*, the number of RNA coding genes is 43.

In a different study, comparative genomics by means of computational methods showed that a total of 462 protein-coding genes are shared among

[9] Mushegian, A.R. and Koonin, E.V., 'A minimal gene set for cellular life derived by comparison of complete bacterial genomes', *Proc. Natl. Acad. Sci. USA* 93, 10268-10273, 1996.

[10] Ortologues are homologous genes in different species that originate from the same ancestral gene in the last common ancestor of the species compared.

three aphid endosymbionts of the genus *Buchnera*.[11] They include genes to synthesize amino acids required by the host and genes necessary for cell division, replication, transcription and protein synthesis. When *B. floridanus* and *W. glossinidia* were added to this analysis, the number of conserved genes in the five endosymbionts decreased to 276. Of these, 156 are also conserved among host-dependent parasites such as *Rickettsia prowazekii*, *Chlamydia trachomatis* and *M. genitalium*. In addition to the 276 protein-coding genes, the five endosymbionts also share 36 RNA specifying genes,[12] which have to be considered in a minimal genome. On the other hand, the number of genes shared between the group of five insect endosymbionts and *Rickettsia*, *Chlamydia* and *Mycoplasma* are 220, 218 and 179, respectively. It is tempting to suggest this last figure as the basic subset of genes required for bacterial cell life. The rest of the genes shared by the five endosymbionts but absent in *M. genitalium* could be involved in endosymbiotic functions.

One of the most thorough studies on minimal genomes has been conducted by Gil and collaborators.[13] Although basically theoretical, this analysis includes data from experimental work and computational comparisons carried out by several groups. Therefore, this contribution by Gil *et al.* could be considered a typical exercise of prediction. These authors defined functions performed in any living cell and then listed the genes that would be necessary to support such functions. The resulting gene list was corrected to fill the gaps in biochemical pathways considered to be essential to maintain homeostasis in any living cell. The hypothetical core able to sustain a functional bacterial cell under ideal conditions turned out to contain 206 protein-coding genes. The number of genes and their respective functions were: 16 implicated in DNA replication and repair, 106 involved in gene transcription and protein synthesis, 15 related to protein folding and secretion, 56 participating in energetic and intermediary metabolism, 4 involved in transport, 1 in cell division and 8 in poorly characterized functions. The latter were identified as essential in *B. subtilis* and therefore added to the list. The authors cautiously acknowledged that there is no conceptual or

[11] Klasson, L. and Anderson, S.G.E., 'Evolution of minimal-gene-sets in host-dependent bacteria', *Trends in Microbiol.* 12, 37-43, 2004.

[12] Gil, R., Silva, F.J., *et al.*, 'The genome sequence of *Blochmannia floridanus*: Comparative analysis of reduced genomes', *Proc. Natl. Acad. Sci. USA* 100, 9388-9393, 2003.

[13] Gil, R., Silva, F.J., Peretó, J. and Moya, A., 'Determination of the core of a minimal bacterial gene set', *Microbiol. Molec. Biol. Rev*. 68, 518-537, 2004.

experimental support for the existence of one particular type of minimal cell. In a sense, there is no single type of minimal metabolism. Therefore, they proposed that their conclusions should be regarded as provisional.

Genome Reduction in the Laboratory

Several experimental approaches permit identification of non-essential genes. These include site-directed gene knockout, global transposon mutagenesis, plasmid insertion mutagenesis and use of antisense RNA.

In a pioneering work,[14] Itaya inserted an antibiotic resistance gene cassette at seventy nine randomly selected chromasomal loci in *Bacillus subtilis*, most of them rare restriction enzyme sites. Only six of the insertions affected bacterial growth in rich medium. Although it is likely that functional redundancy may lead to an underestimation of essential genes, the author also tested multiple (7-, 12- and 33) fold mutations among identified dispensable loci. These highly mutated strains retained the ability to form colonies. The indispensable DNA size was calculated by satistical analysis to be in the range of 318-562 kb. Considering that bacterial open reading frames encompass about 1 kb, this minimal genome would harbor between 300 and 500 genes.

Transposons are segments of DNA that can move from one location in a genome to another, often disrupting gene function by insertion. Once the location of enough transposon insertions are defined by DNA sequencing, the researcher is able to deduce with some certainty that regions in which transposon insertions are not observed are likely to be essential for viability. In other words, transposition not affecting cell viability allow the identification of non-essential genes. Global transposon mutagenesis has been used to identify non-essential genes in *M. genitalium*. In a first attempt,[15] 685 insertion events led to the conclusion that 265 to 350 of the 482 protein-coding genes are absolutely essential for growth under laboratory conditions. Of these, 100 are of unknown function. To confirm that the insertions had been indeed disruptive, only those found after the first three codons and before the 3'-most 20% of the coding sequence were considered in this work.

[14] Itaya, M., 'An estimation of minimal genome size required for life', *FEBS Lett*. 362, 257-260, 1995.

[15] Hutchinson, C.A., Peterson, S.N., Gill, S.R., Cline, R.T., White, O., Fraser, C.M., Smith, H.O. and Venter, J.C., 'Global transposon mutagenesis and a minimal Mycoplasma genome', *Science* 286, 2165-2169, 1999.

In a more recent report, the same group expands this work by proving gene dispensability after isolation and characterization of pure colonies, a precaution that they did not originally take.[16] This step is necessary because there is always the possibility that other cells in the same pool of mutants may supply a gene product. In this refined study, the authors found that 100 instead of 120 genes are nonessential. None of the genes suspected to be essential for growth (DNA replication, glycolisis, cytoskeleton, etc.) were disrupted, including the 43 RNA-coding genes. This new study also showed that in spite of its small genome, *M. genitalium* possesses some enzymatic redundancy, a property that would mask the requirement of a gene disrupted by insertion mutagenesis. Taking this into consideration, 5 genes were added to the minimal set, thus ending up with a total of 387 essential protein-coding genes. This number is higher than the same group's initial prediction and surprisingly larger than those predicted in the theoretical calculations described above. Another unexpected characteristic of this essential gene set is that it includes 110 proteins of unknown function. On the other hand, it is likely that several of the dispensable genes are involved in the maintenance of *M. genitalium* in the human urogenital tract, its natural habitat.

Genome-scale transposon mutagenesis has also been conducted with the bacterium *H. influenzae*. The number of putative essential genes in this case turned out to be 670, also a much higher figure than the one deduced by theoretical comparison with *Mycoplasma*.[17] This implies that about 40% of the genome is essential under the conditions tested. Some of the genes identified have proven dispensable in other bacteria, whereas as many as 259 genes lack a defined functional role. It is possible that transposon mutagenesis overestimates the size of the minimal gene set by misclassification of non-essential genes that slow down growth without arresting it. Moreover, in this particular study, the authors included in the minimal set genes that had been the target of single insertions (191), following the rationale that some essential genes encode non-essential domains. On the other hand, computation can underestimate the set because it takes into account only those genes that have remained similar enough during evolution to be considered as canonical orthologues.

[16] Glass, J.I., Assad-García, N., Alperovich, N., Yooseph, S., Lewis, M.R., Maruf, M., Hutchinson, C.A., Smith, H.O. and Venter, J.C., 'Essential genes of a minimal bacterium', *Proc. Natl. Acad. Sci.* USA 103, 425-430, 2006.

[17] Akerley, B.J., Rubin, E.J., Novick, V.L., Amaya, K., Judson, N. and Mekalanos, J.J., 'A genome-scale analysis for identification of genes required for growth or survival of *Haemophilus influenzae*', *Proc. Natl. Acad. Sci. USA* 99, 966-971, 2002.

Another study utilizing a similar experimental technique known as plasmid insertion mutagenesis was conducted with the Gram-positive bacterium *Bacillus subtilis*. Unexpectedly, only 271 genes out of the 4,100 comprising the genome appeared essential when they are individually inactivated.[18] However, this approach does not detect essential functions encoded by redundant genes because a single gene is inactivated in each mutant strain. Therefore, the minimal set mentioned before may be an underestimation. The great majority of the 271 genes encode functions related with information processing, cell envelope, shape, division and energetics. Surprisingly, the authors found that genes encoding glycolytic enzymes are required for growth even though the experiments were conducted in a rich medium that contains numerous compounds that could provide energy under aerobic conditions. This suggests that these enzymes may have unknown functions in the cell. About 50% of the genes of this minimal set are found in all bacteria, even in those with the smallest genomes.

As mentioned previously, interference of gene expression with antisense RNA constitutes another experimental approach to identify essential genes. In this case, gene function is inhibited by formation of a duplex RNA structure between the target mRNA and the antisense RNA introduced to the cell. Essential genes are identified after conditionally expressing random genomic fragments representing the entire genome linked to an inducible promoter, and then screening for those cloned fragments whose expression blocks growth. DNA sequencing and BLAST analysis against the annotated genome identify the genes targeted by antisense RNA. Forsyth *et al.* followed this strategy with *Staphylococcus aureus*, the most frequent causative agent of nosocomial infections in humans.[19] These authors found that out of the 2,595 protein coding genes in this microorganism, 658 are essential for growth. Homology comparison showed that 168 of these genes are found in the genome of *M. genitalium*. Interestingly, a similar approach followed by another group with the same bacterium showed that antisense inactivation of only 150 genes led to lethal or growth inhibitory effects.[20] The reason for this discrepancy is not known.

[18] Kobayashi, K., Ehrlich, S.D., *et al.*, 'Essential *Bacillus subtilis* genes', *Proc. Natl. Acad. Sci. USA* 100, 4678-4683, 2003.

[19] Forsyth, R.A., Haselbeck, R.J. *et al.*, 'A genome-wide strategy for the identification of essential genes in *Staphylococcus aureus*', *Mol. Microbiol.* 43, 1387-1400, 2002.

[20] Ji, Y., Zhang, B., Van Horn, S.F., Warren, P., Woodnutt, G., Burnham, M.K.R. and Rosenberg, M., 'Identification of critical staphylococcal genes using conditional phenotypes generated by antisense RNA', *Science* 293, 2266-2269, 2001.

Comparison of Both Approaches and Concluding Remarks

Both theoretical and practical approaches for calculating minimal gene set numbers have some pitfalls.[21] First, computational methods rely on our ability to identify actual orthologues in species that are distantly related, as well as on filling some essential functions corresponding to NOGD. In spite of the general belief that we have identified the essential functions in a cell, it is expected that some cases of NOGD may remain hidden. In addition, different solutions may have evolved for each function in different organisms, giving rise to vast combinatorial possibilities. For these reasons, a minimal gene-set derived by computational comparison of genomes is probably an underestimate.

On the other hand, for technical reasons, the mutagenizing protocols described above are able to target only about 50% the genes, although this is considered statistically sufficient for reliable extrapolations. The knockout experiments also tend to score as essential genes that slow down microbial growth but do not abolish it, leading to a potential overestimation of the minimal gene-set. This possibility is partially compensated by the fact that knocking out of individual genes does not unveil the synthetic lethal genes. That is, those for which only simultaneous mutation results in a lethal phenotype.

In spite of these drawbacks, the number of essential genes reached by the computational and experimental approaches are in the same range. Moreover, the protein functions encoded in both minimal sets are strikingly similar, with a notorious enrichment of genes involved in the processing of genetic information, a rather low number of genes encoding metabolic enzymes and very few genes of unknown function.

Are further reductions possible? The work by Gil *et al.* described above provides copious arguments supporting a minimal genome of 206 genes. However, the functions considered by these authors are those known in extant cells. It is quite obvious that primitive cells must have been far more simple. For example, the protein synthesizing machinery may have contained less ribosomal proteins, or perhaps, proteins (enzymes) were more commonly multifunctional and therefore less specific. Today it is widely accepted that increased cell complexity is due to gene duplication. Indepen-

[21] Koonin, E.V., 'Comparative genomics: minimal gene-sets and the last universal common ancestor', *Nature Rev. Microbiol.* 1, 127-136, 2003.

dent mutations in the resulting paralogue[22] genes gradually results in the addition of more functions in the cell. As the matter of fact, gene duplication together with horizontal gene transfer are considered the main sources of microbial diversity in the biosphere. Therefore, analysis of the genes comprising a minimal genome may give information with respect to the key functions in primitive cells, since the simultaneous presence of orthologues in the three primary kingdoms of life is a sign of a function already present in a common ancestor.

De Novo *Synthesis of a Minimal Genome*

Once a minimal genome suitable for a specific set of environmental conditions is defined with a certain degree of confidence, an inevitable question emerges: could a functional genome be assembled that would allow metabolic activity and replication? This 'bottom up' approach, as opposed to the 'top down' strategy of knocking out genes of an existing microbe, entails an enormous task. The means for synthesizing small pieces of DNA are presently available, but assembling an entire genome including the regulatory sequences represents a challenge of considerable magnitude. This is mainly due to the possibility of contamination with truncated species and the introduction of typographical mistakes into the code. Interestingly, construction of synthetic genomes has been accomplished with small viral chromosomes such as poliovirus (7,440 bases)[23] and bacteriophage ϕX174,[24] whose infectious genome possesses 5,386 bp. The latter involved the use of new methods that significantly improved the speed and accuracy of genomic synthesis.

Craig Venter and Hamilton Smith are determined to construct a microorganism with a synthetic minimal genome. These same investigators conducted the studies of transposon mutagenesis with *Mycoplasma* and the synthesis of ϕX174.[25] Their approach involves selection of genes comprising

[22] Paralogs are genes in the same cell that derive from a single ancestral gene.

[23] Cello, J., Paul, A.V. and Wimmer, E., 'Chemical synthesis of poliovirus cDNA: Generation of infectious virus in the absence of natural template', *Science* 297, 1016-1018, 2002.

[24] Smith, H.O., Hutchinson, C.A., Pfannkoch, C. and Venter, J.C., 'Generating a synthetic genome by whole genome assembly: ϕX174 bacteriophage from synthetic oligonucleotides', *Proc. Natl. Acad. Sci. USA* 100, 15440-15445, 2003.

[25] Needless to say, Venter had a leading role in the sequencing of the human genome and has thereafter conducted interesting metagenomic studies.

the minimal genome, chemical synthesis of all genes, stitching in some undetermined fashion and introduction into a bacterial cell whose own genome has been irreversibly damaged or destroyed. But Venter's ambition doesn't stop there. His goal is to use this minimal microorganism as a foundation for building cells harboring additional genes that would enable them to consume pollutants from the environment or to produce hydrogen fuel at an industrial scale. In 2003, he predicted that he would accomplish the task in three years.[26] Obviously, this has not yet happened, or at least, it has not been reported. Being aware that this project raises some ethical concerns, Venter requested an ethics committee from Stanford University to weigh the risks of creating new life forms. The panel acknowledged that there is a large technological gap between defining a minimal gene set and actually 'creating life', ruling out moral obstacles to assemble a new microorganism. However, it recommended responsible use of the new technology, since it could pose threats to public health and safety due to possible environmental contamination or the development of biological weapons.

As mentioned above, predictability in science allows us to foretell a given situation or event based on observation and reasoning. Can we predict whether scientists will ever create microbial life in the laboratory? This seems to be an extremely difficult prediction to make. Some years ago, this goal would have been considered science fiction. However, with the rapid development of novel and highly sophisticated technologies, perhaps no one would be in a position to predict that life will never by created *de novo* in the laboratory.

[26] Zimmer, C., 'Tinker Taylor: Can Venter stitch together a genome from scratch?', *Science* 299, 1006-1007, 2003.

THE NEW POSSIBILITIES OF PREDICTION AND PREVENTION OF CANCER

UMBERTO VERONESI

Cancer is an environmental disease. Carcinogenic agents are extensively spread throughout the environment, inclusive of all the elements in which we live. Therefore all agents, of any type, which come in contact with the human body, are defined environmental agents. They include carcinogens in the air, in the water and in food.

In addition there are all those agents which form part of our daily behaviour, cigarette smoke, cosmetics of any type, pharmaceutical drugs, children's toys coated with substances which may contain chemical carcinogens, pesticides, asbestos, benzene in gasoline and in addition radiation of different kinds, such as ionizing radiation from the earth, from cosmic rays, and from radon gas present in our homes. There are viruses in the environment, such as HPV for cancer of the uterus, bacteria, helicobacter pylorus for gastric cancer and parasites such as schistosoma haematobium for cancer of the bladder.

All those agents, although innumerable and varied, have a common mechanism to induce cancer. They damage the DNA of a somatic cell creating a lesion, which we call a 'mutation', of one or more genes. This lies at the root of the carcinogenic process, as the specific gene mutation gives the cell the ability to proliferate without the usual limitations imposed by the harmony of all our organs and tissues. In other words the mutated cell is a cell which escapes from the normal 'program' of the organism and becomes a deviated cell.

Therefore this minimal error in the complex 'software' which regulates all our functions may lead to serious and tragic consequences.

Fortunately, our body has a good 'repair system', rich in enzymes which are able to repair the damage of the DNA and eliminate the dangerous mutation.

If the DNA repair enzymes work properly there will be no cancer; if the enzymes are not active enough the mutated cell will progressively develop into a true carcinoma (Fig. 1). The carcinogenic process generally takes a certain number of years, from 3 to 20.

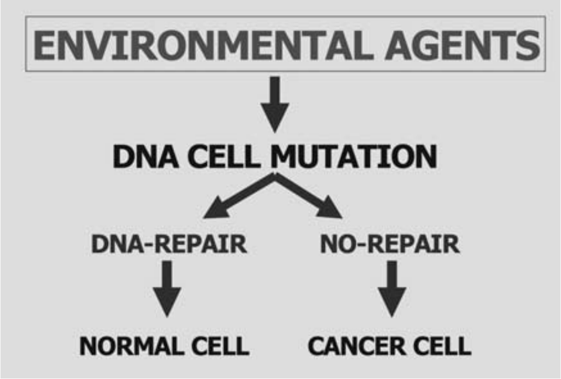


Figure 1.

The natural defence mechanism via DNA repair enzymes is genetically regulated and this accounts for the personal or familial predisposition to develop cancer lesions.

Constitutional, genetic	3%
Constitutional acquired (reproductive endocrine factors)	12%
Environmental (chemical, physical viral)	85%

Figure 2. Factors which lie at the origin of cancer.

The causes of cancer are therefore mainly environmental. True hereditary cancers are uncommon, while a number of tumours, in particular breast carcinoma, are linked to reproductive endocrine conditions.

How did we discover that most cancers have an environmental origin? The answer came from epidemiological research in three different areas. The first was the different incidence of the various tumours in different countries (geographic pathology). Figure 3 (see page 274) shows the incidence of the most common types of tumour in the various parts of the world. The fact that there is a great difference among various countries is a reflection of the different lifestyles of the world populations. Therefore lung cancer is common among populations where cigarette smoking is common, cancer of the uterus in countries where HPV infection is endemic, cancer of the liver in areas where hepatitis B virus is diffuse and so on (Figs. 4-12, see pages 274-275).

The second area of epidemiological research refers to migrant population studies. In fact, if cancer really is genetically determined, when a population living in a country where there is a high incidence of a certain cancer decides to emigrate to a country where the incidence of that particular cancer is low, then the migrant's risk should remain at the same level. The opposite will hold if the cancer is due to environmental factors. All the studies are in favour of the latter hypothesis.

In fact, when Japanese people leave Japan, where gastric cancer incidence is the highest in the world, to migrate to the United States, where gastric cancer incidence is very low, in a couple of generations the incidence of gastric cancer diminishes to reach the levels of the American populations. The third type of study is time-trend analysis of cancer over a period of decades. For example in 1950s Italy women were affected by cancer of the uterus and men by cancer of the stomach. At the end of the century both uterus and gastric cancer were no longer the leading types of cancer while women were affected mainly by cancer of the breast and men by cancer of the lung. The latter is undoubtedly largely due to the habit of smoking and the former to the fact that women have very few pregnancies, or little or no breast feeding and above all have their first child much later in life than they would have had 50 years ago.

How then, can mortality rates be reduced?

Mortality may be reduced by (a) reduction of carcinogens in the environment, (b) use of drugs or active principles which will block the carcinogenic process and (c) an early detection of cancer and adequate, timely treatment.

1. Reduction of Carcinogens in the Environment

In 1985 two renowned British epidemiologists Doll and Peto, conducted a review of a large number of papers addressing cancer incidence, and prepared a list of categories of carcinogenic agents from which it could be concluded that it is in the food that we should look if we want to find the key carcinogens. Tobacco smoking is the second cause, according to Doll and Peto, followed by infectious agents and reproductive factors. Air pollution only accounts for a mere 2% (Table 1). Food by itself may stimulate uncontrolled cell proliferation but, above all, it is a carrier of many carcinogenic agents present in the environment. Meat for example comes from animals which graze in the open air and their very act of eating the grass introduces all the carcinogenic agents polluting the ground into their bodies.

TABLE 1. CANCER CAUSES

Kind of exposition	Risk distribution
Food	35
Tobacco	30
Viral	10
Reproductive Factors	7
Occupational Activities	4
Geographic Factors	3
Air Pollution	2
Medical Drugs	1
Unknown Factors	?

From Doll and Peto 1985.

Ingesting a lot of meat, in fact, increases the risk of intestinal cancer. Cancers are therefore closely related to nutrition and diet, with an increased risk arising from the eating highly caloric food rich in animal fat and a reduced risk arising from the consumption of fruit and vegetables. Figure 13 (page 276) shows the strict correlation between meat consumption and colorectal cancer incidence in the five continents and Figure 14 (page 276) shows how the increased consumption of fruit and vegetables leads to a notable reduction of cancer risk in various organs.

Furthermore, food may be contaminated by residual pesticides and particularly one known as aflatoxin which is one of nature's most active carcinogenic agents. Aflatoxins are a group of secondary metabolites which are cancer causing by-products of a mould that grows on grains and nuts, particularly peanuts.

Although aflatoxin is most commonly produced when the potentially affected foods are incorrectly stored, recent studies have documented that it arises in fields, particularly as a consequence of severe climatic changes or if the plants undergo attacks by insects. Most industrialized nations impose strict regulations on aflatoxin levels in food for human consumption. However, many of these products are employed in animal feed, and if an animal consumes infected food, the aflatoxin is transmitted to humans via contaminated milk and meat products. Aflatoxin is a carcinogen for certain animals, particularly cattle.

Among humans, it is associated with liver cancer, particularly in Third World nations where malnutrition and other health problems are also prevalent.

Animals grazing on bracken may exhibit various signs of toxicity, including tumours in the upper gastrointestinal tract and bladder, which are attributable to the carcinogen ptaquiloside. The corresponding glucoside may be present in bracken at a concentration of 13,000 ppm. Metabolism of this compound gives rise to alkylation adducts in DNA.

Milk from bracken fern-fed cows induces cancer in experimental animals. Bracken may pose a carcinogenic hazard for humans in populations identified as exposed in Japan, Costa Rica and the United Kingdom.

There are some organochlorines, such as DDT and other pesticides, which are resistant to degradation, and are highly lipid-soluble. They thus persist in the environment and are bioconcentrated in the human food chain. Related industrial chemicals such as polychlorinated biphenyls behave in the same manner. DDT and a number of other organochlorine pesticides cause liver cancer in rats. DDT has been especially linked with an increased risk of pancreatic cancer, breast cancer, lymphoma and leukaemia in humans.

Some organochlorines exhibit sex steroid activity in relevant assay systems, and these pesticides are considered to potentially subvert endocrine-regulated homeostasis.

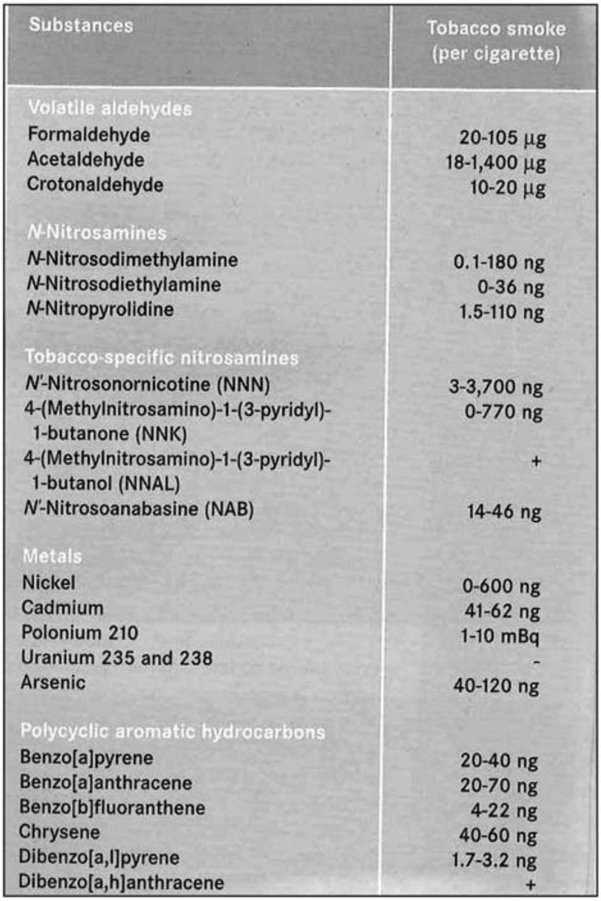
When meat and fish are cooked at high temperatures, certain heterocyclic amines are formed as a result of the pyrolysis of two amino acids, namely creatine and creatinine. Heterocyclic amines are carcinogenic in various organs of mice, rats and non-human primates, although their carcinogenic potential in humans is yet to be established.

Various genetic polymorphisms mean that heterocyclic amine metabolism can vary from individual to individual.

Tobacco

The smoking of tobacco is known to be the main cause of human cancer-related deaths worldwide. Smoking most commonly causes lung cancer. For a smoker, lung cancer risk is related to tobacco smoking parameters in accordance with the basic principles of chemical carcinogenesis: carcinogen dose, the duration of administration and exposure intensity are known to be risk determinants. Here, women are at least as susceptible as men. It is consistently clear that there is an increased risk of lung cancer (relative to a non-smoker) at the lowest level of daily consumption, and this bears at least a linear relation to increasing consumption. The risk is also proportional to the duration of smoking. Hence, the annual death rate from lung cancer among 55-64 year-olds who smoke 21-39 cigarettes per day is about three times that of those who commenced smoking at the age of 15 than it is for those who started smoking at 25. Smoking of black tobacco cigarettes represents a greater risk for most tobacco-related cancers than does smoking of blond cigarettes. Similarly, filtered and low-tar cigarettes entail a lower risk for most tobacco-related cancers than do unfiltered and high-tar cigarettes. Looking at communities worldwide, incidence of lung cancer varies dramatically. High rates are observed in parts of North America, while developing countries have the lowest rates. In the USA, Europe and Japan, 83-92% of lung cancers in men and 57-80% of lung cancers in women are tobacco – related. In addition to lung cancer, smoking causes cancers of the larynx, oral cavity, pharynx, oesophagus, pancreas, kidney and bladder. Dose-response relationships between number of cigarettes smoked and risks for developing these cancers have been consistently found. A pattern of decreased risk of lung and other smoking dependent cancers is commonly observed to follows smoking cessation ('quitting') relative to those who continue to smoking.

The relative risk of cancer at most sites is markedly lower than that of current smokers after five years' cessation, although risks for bladder cancer and adenocarcinoma of the kidney appear to persist longer before falling off. Despite the clearly established benefits of cessation, the risk for ex-smokers does not decrease to that enjoyed by so-called 'never smokers'. Exposure to environmental tobacco smoke causes lung cancer and possibly laryngeal cancer. Although the burden of disease is much less than in active smokers; the relative risk has been estimated at about 1.15-1.2 (Fig. 15).



Substances	Tobacco smoke (per cigarette)
Volatile aldehydes	
Formaldehyde	20-105 μg
Acetaldehyde	18-1,400 μg
Crotonaldehyde	10-20 μg
N-Nitrosamines	
N-Nitrosodimethylamine	0.1-180 ng
N-Nitrosodiethylamine	0-36 ng
N-Nitropyrolidine	1.5-110 ng
Tobacco-specific nitrosamines	
N'-Nitrosonornicotine (NNN)	3-3,700 ng
4-(Methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK)	0-770 ng
4-(Methylnitrosamino)-1-(3-pyridyl)-1-butanol (NNAL)	+
N'-Nitrosoanabasine (NAB)	14-46 ng
Metals	
Nickel	0-600 ng
Cadmium	41-62 ng
Polonium 210	1-10 mBq
Uranium 235 and 238	-
Arsenic	40-120 ng
Polycyclic aromatic hydrocarbons	
Benzo[a]pyrene	20-40 ng
Benzo[a]anthracene	20-70 ng
Benzo[b]fluoranthene	4-22 ng
Chrysene	40-60 ng
Dibenzo[a,l]pyrene	1.7-3.2 ng
Dibenzo[a,h]anthracene	+

Figure 15.

Infectious Agents

For more than 100 years it has been known that infectious agents can cause cancer. In 1911 Peyton Rous demonstrated that sarcomas in chickens were caused by an infectious agent, later identified to be a virus. Today, there is a raft of evidence, experimental and epidemiological, to indicate that a variety of infectious agents constitutes one of the main causes of cancer worldwide. Viruses are the main agents, at least eight different viruses being associated with particular tumour types, with varying degrees of certainty. Some 2,000 million people worldwide have serological evidence of current or past hepatitis B virus (HBV) infections and about 350 million of these people are chronic carriers of the virus. It has been estimated that 60% of primary liver cancer cases worldwide and 67% of cases in developing countries are attributable to chronic persistent infection with HBV.

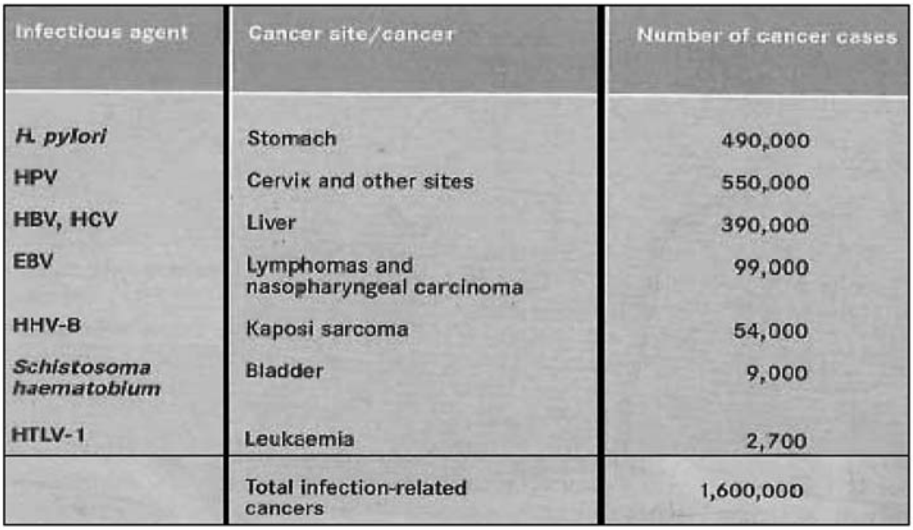
Approximately 25% of liver cancer cases worldwide can be attributed to HCV.

HPV DNA is found in virtually all invasive cervical cancers, indicating that HPV is an essential cause.

Epstein-Barr virus (EBV) infection is ubiquitous. In developing countries, infection is acquired in childhood, while in developed countries infection is delayed until adolescence.

Individuals with high titres of antibodies to various early and late EB antigens are at a higher risk of developing Burkitt's lymphoma and Hodgkin's disease.

Human T-cell lymphotropic virus (HTLV-1) infection occurs in clusters in Japan, Africa, the Caribbean, Colombia and Melanesia. There may be as many as 20 million people worldwide infected with this virus. A strong geographical correlation suggests that HTLV-1 is the main etiological factor in adult Tcell leukaemia/lymphoma. Human herpesvirus 8



Infectious agent	Cancer site/cancer	Number of cancer cases
H. pylori	Stomach	490,000
HPV	Cervix and other sites	550,000
HBV, HCV	Liver	390,000
EBV	Lymphomas and nasopharyngeal carcinoma	99,000
HHV-B	Kaposi sarcoma	54,000
Schistosoma haematobium	Bladder	9,000
HTLV-1	Leukaemia	2,700
	Total infection-related cancers	1,600,000

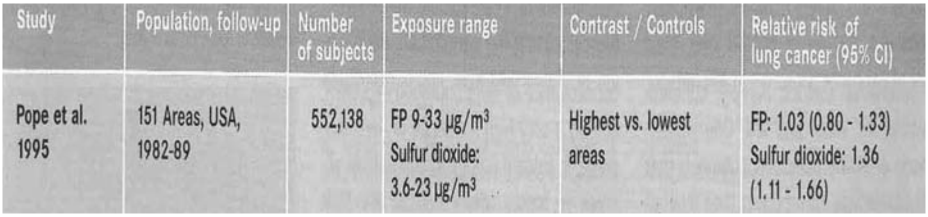
Figure 16.

(HHV-8) infection appears to be common in Africa and in some Mediterranean countries but rare elsewhere. HHV-9 DNA has been detected in over 90% of Kaposi sarcomas and rarely in control patients. Helicobacter pylori infection is one of the world's most common bacterial infections. H. pylori undoubtedly plays a role in gastric cancer, but there are other contributory cofactors, such as diet (Fig. 16).

Environmental Pollution

Air, water and soil pollution is estimated to account for 1-4% of all cancers. A small proportion of lung cancers (>5%) can be ascribed to industrial effluent, engine exhaust output and other outdoor toxins. Carcinogenic indoor air pollutants include tobacco smoke, and cooking fumes in particular regions, including parts of Asia.

The carcinogenic pollutants for which most information is available include toxic asbestos in urban air, indoor air pollutants, chlorination by products and other contaminants of drinking water. The carcinogenic hazard of asbestos dust has been recognized since the 1950s (Fig. 17).



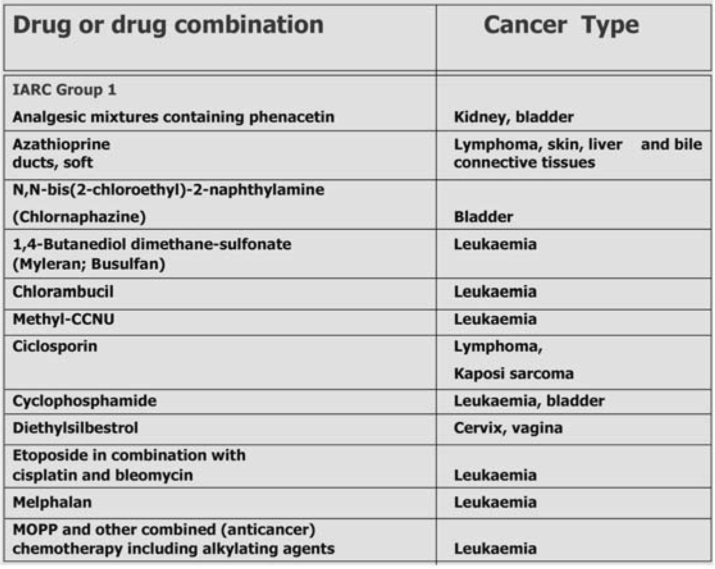
Study	Population, follow-up	Number of subjects	Exposure range	Contrast / Controls	Relative risk of lung cancer (95% CI)
Pope et al. 1995	151 Areas, USA, 1982-89	552,138	FP 9-33 µg/m³ Sulfur dioxide: 3.6-23 µg/m³	Highest vs. lowest areas	FP: 1.03 (0.80 - 1.33) Sulfur dioxide: 1.36 (1.11 - 1.66)

Figure 17.

Medicinal Drugs

Certain cancer treatment drugs may, on rare occasions, lead to second primary tumours.

Modern medicine has hundreds of drugs at its disposal, many of which are an essential part of the doctor's armamentarium in effectively treating a vast panoply of diseases. A small fraction of such drugs have been found, however, to have the side-effect of carcinogenicity. This occurs mostly in certain drugs that have to be administered at high doses or for prolonged periods of time (Fig. 18).



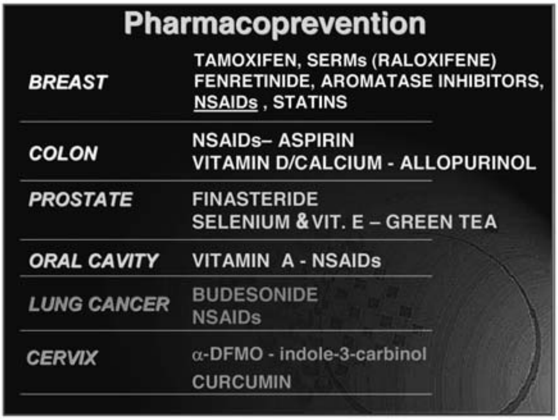
Drug or drug combination	Cancer Type
IARC Group 1 Analgesic mixtures containing phenacetin	Kidney, bladder
Azathioprine ducts, soft	Lymphoma, skin, liver and bile connective tissues
N,N-bis(2-chloroethyl)-2-naphthylamine (Chlornaphazine)	Bladder
1,4-Butanediol dimethane-sulfonate (Myleran; Busulfan)	Leukaemia
Chlorambucil	Leukaemia
Methyl-CCNU	Leukaemia
Ciclosporin	Lymphoma, Kaposi sarcoma
Cyclophosphamide	Leukaemia, bladder
Diethylsilbestrol	Cervix, vagina
Etoposide in combination with cisplatin and bleomycin	Leukaemia
Melphalan	Leukaemia
MOPP and other combined (anticancer) chemotherapy including alkylating agents	Leukaemia

Figure 18.

2. Pharmacoprevention

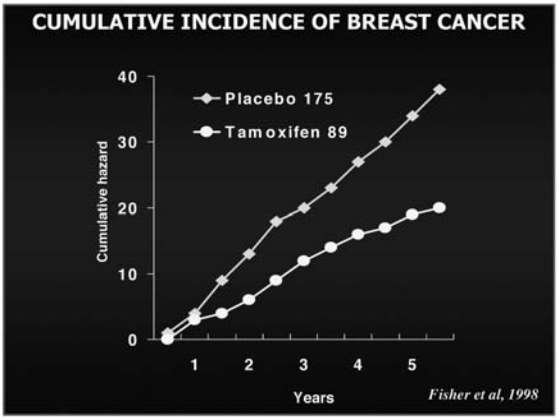
The pharmacological prevention of cancer represents a comparatively novel field in clinical oncology, but it offers a very promising approach to reducing the burden of cancer and its incidence. Other medical disciplines, such as cardiology have taken this route, whereby it is common practice to treat subjects at higher risk for cardiovascular disease long before clinical evidence. This has made a definite contribution to a lower mortality. A similar strategy can be adopted for cancer prevention in ‘at higher risk’ subjects.

The peculiarity of carcinogenesis is that it is a multistep, multipath and multifocal process, involving a series of genetic and epigenetic alterations which develop from genomic instability all the way to the final development of cancer. This is the key notion lying behind the rationale for intervention in the initial steps of the process, by employing natural or synthetic agents potentially able to delay, arrest or even reverse the pathogenesis of cancer.



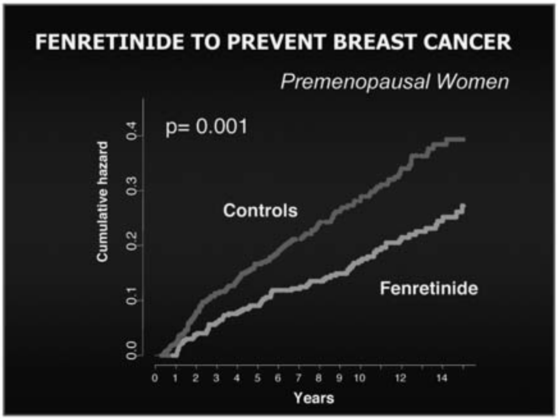
Pharmacoprevention

BREAST	TAMOXIFEN, SERMs (RALOXIFENE) FENRETINIDE, AROMATASE INHIBITORS, NSAIDs , STATINS
COLON	NSAIDs– ASPIRIN VITAMIN D/CALCIUM - ALLOPURINOL
PROSTATE	FINASTERIDE SELENIUM & VIT. E – GREEN TEA
ORAL CAVITY	VITAMIN A - NSAIDs
LUNG CANCER	BUDESONIDE NSAIDs
CERVIX	α-DFMO - indole-3-carbinol CURCUMIN



CUMULATIVE INCIDENCE OF BREAST CANCER

Fisher et al, 1998



FENRETINIDE TO PREVENT BREAST CANCER

Premenopausal Women

Since the process is mostly very long (10-20 years, sometimes more), there is potentially a great deal of time to assess the true risk and intervene with nutrients and/or pharmacological agents which will interrupt the chain of molecular events long before the onset of clinical symptoms. This may prove of particular use where solid tumours are concerned, which are often characterised by multifocality and metachronous growth of lesions resulting from the plausible concept of field carcinogenesis and intraepithelial clonal spread.

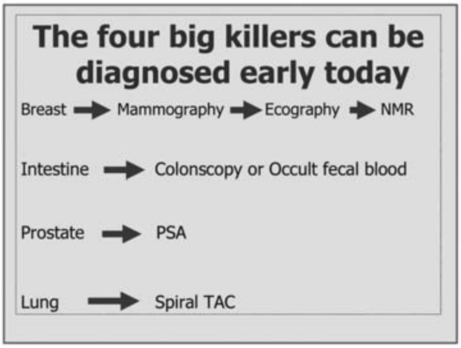
Recently, a number of compounds have shown to be clinically effective at various organ levels, covering all the three settings in which prevention may be typically divided into, namely: primary, where the goal is to prevent the onset of the disease, selecting healthy cohorts at high risk because of their environmental or lifestyle or familial/genetic factors; secondary, aimed at treating a population possessing a premalignant condition or an in situ neoplasia thereby blocking its evolution to cancer; and tertiary, which is aimed at protecting against second primary tumours in subjects previously cured for a cancer.

3. Early Detection

It is an oft-quoted truth in cancer care that if the tumour is discovered at an early stage – the chances of cure are much higher compared with cases where diagnosis is late. A case in point is the near-epidemic increase in breast cancer incidence being due to the introduction of population-based mammography screening. The analysis of large randomized trials has shown that in women aged 50 to 69 years, mammography screening can reduce mortality from breast cancer by 25-30%. For women in the 40-49 year age group the screening efficacy is significantly less. Considering other cancers, early detection programs are thriving for cancers of the uterine cervix, intestine, prostate and lung (Fig. 19).

With a more widespread use of screening programs and educational initiatives, future cancer mortality for the more common tumours is expected to see a notable reduction.

One of the greatest hopes lies in the technological development of diagnostic tools. So-called diagnostic imaging equipment is becoming more and more sophisticated and will form the basis for new progress in early cancer detection (Figs. 20-24, see page 277).



The four big killers can be diagnosed early today

Breast → Mammography → Ecography → NMR

Intestine → Colonscopy or Occult fecal blood

Prostate → PSA

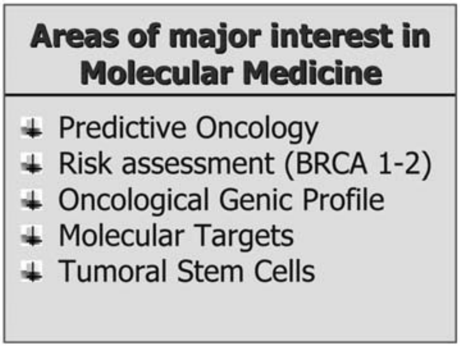
Lung → Spiral TAC

Figure 19.

4. Molecular Oncology

The discovery of the sequences of human DNA has led to the *'decodification'* of the role of the 30.000 genes present in every human cell and given rise to biomolecular oncology, the main medical revolution of recent decades. The fields of interest are many and are of great value (Fig. 25) and will in a decade lead to a number of discoveries which may herald the victory of our fight against cancer.

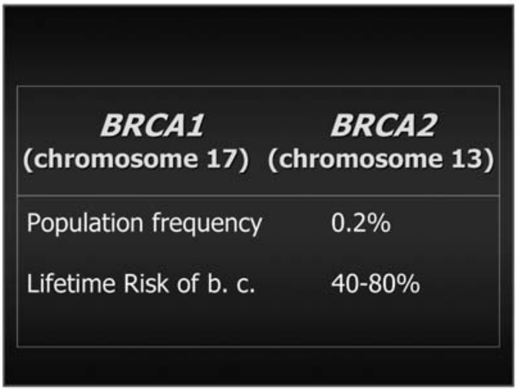
In conclusion, although the individual assessment of the risk of developing cancer may be an arduous task, prevention of cancer is possible, provided that the population is correctly informed and that the public health authorities are aware of, and sensitive to, the issues involved (Figs. 26-27).



Areas of major interest in Molecular Medicine

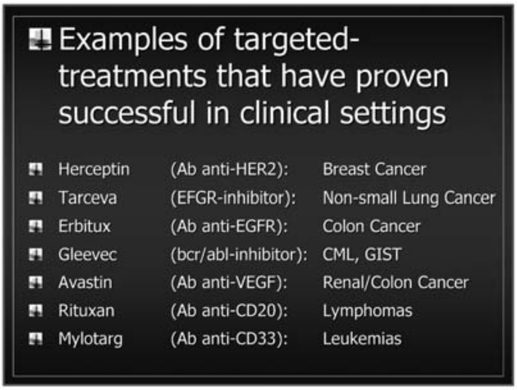
- Predictive Oncology
- Risk assessment (BRCA 1-2)
- Oncological Genic Profile
- Molecular Targets
- Tumoral Stem Cells

Figure 25.



BRCA1 (chromosome 17)	*BRCA2* (chromosome 13)
Population frequency	0.2%
Lifetime Risk of b. c.	40-80%

Figure 26. The mutations in chromosome 17 (BRCA1) AND 13 (BRCA2) create a condition of notably increased risk for breast cancer and, to a more limited extent, for ovarian carcinoma.



Examples of targeted-treatments that have proven successful in clinical settings

Herceptin	(Ab anti-HER2):	Breast Cancer
Tarceva	(EFGR-inhibitor):	Non-small Lung Cancer
Erbitux	(Ab anti-EGFR):	Colon Cancer
Gleevec	(bcr/abl-inhibitor):	CML, GIST
Avastin	(Ab anti-VEGF):	Renal/Colon Cancer
Rituxan	(Ab anti-CD20):	Lymphomas
Mylotarg	(Ab anti-CD33):	Leukemias

Figure 27. Molecular oncology studies have led to the discovery of many molecules and monoclonal antibodies which direct themselves towards a specific molecular target due to a specific DNA mutation.

REFERENCES

1. *World Cancer Report*. WHO – OMS – IARC. Edited by Stewart BW and Kleihues P. IARC Press, Lyon 2003.
2. Doll R, Peto J. *Asbestos: effects on health of exposure to asbestos. A report to the Health and Safety Commission*, London: HMSO, 1985.
3. Fisher B, Costantino JP, Wickerham DL, Redmond CK, Kavanah M, Cronin WM, Vogel V, Robidoux A, Dimitrov N, Atkins J, Daly M, Wieand S, Tan-Chiu E, Ford L, Wolmark N. Tamoxifen for prevention of breast cancer: report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study. *J Natl Cancer Inst.* 1998; 90(18):1371-88.
4. Veronesi U, Mariani L, Decensi A, Formelli F, Camerini T, Miceli R, Di Mauro MG, Costa A, Marubini E, Sporn MB, De Palo G. Fifteen-year results of a randomized phase III trial of fenretinide to prevent second breast cancer. *Ann Oncol*. 2006 Jul;17(7):1065-71.
5. Veronesi U, Bonanni B. Chemoprevention: from research to clinical oncology. *Eur J Cancer*. 2005;41(13):1833-41.

STOCHASTIC GENETIC VARIATIONS AND THEIR ROLE IN BIOLOGICAL EVOLUTION

WERNER ARBER

1. Introduction

It is still a widespread view, particularly in the general public, that all genes carried in the genome determine their encoded functions accurately and fully predictably. Scientists, however, realize more and more that this does not correspond to the reality by far. There are many reasons for a higher complexity, such as epigenetic phenomena, protein modifications, regulatory circuits in the control of gene expression and environmental impacts. Studies on the generation of genetic variants, that represent the driving force of biological evolution, have revealed rather unexpected properties of specific gene products that can be called variation generators. These proteins generate genetic variants with a high degree of unpredictability with regard to the site of mutagenesis on the genome and with regard to the time when a variation occurs. We will illustrate this situation with a few examples that were obtained by experimentation in microbial genetics. The evolutionary role of these activities will be discussed in the context of a more general theory of molecular evolution.

2. From the Neodarwinian Theory to a Theory of Molecular Evolution

The so-called modern evolutionary synthesis resulting in the Neodarwinian theory of biological evolution united, around 1940, Darwin's concept of natural selection with the notion of spontaneous mutagenesis as the source of hereditary phenotypic variation (Mayr, 1982). Shortly thereafter, DNA was identified as the carrier of genetic information (Avery *et al.*, 1944) and the filamentous, double helical structure of DNA molecules was described (Watson and Crick, 1953). This then gave rise to molecular genetics. In the meantime it has become known that changes of inheritable traits are linked to alterations of nucleotide sequences in the underlying genetic information. However, not all changes occurring in genomic sequences result in an altered phenotypic trait, for a number of understood reasons. Because of this situation, two different definitions for the synonymously used terms of genetic variation and genetic mutation are found in the genetic literature. In classical genetics, a mutation is an altered phenotype that becomes transmitted to the progeny. In contrast, in molecular, reverse genetics a mutation is an alteration in the nucleotide sequence of the DNA.

In the context of today's knowledge on molecular evolution the process of biological evolution stands on three pillars. One of these is genetic variation that can, to some degree, affect life activities either positively or negatively. The second pillar is natural selection, i.e. the result of the interaction of individual organisms with their encountered environmental constraints. These depend both on physico-chemical properties of the environment and on the activities of all other organisms living in the same ecological niche. The third pillar is isolation that can be reproductive or geographic.

The generation of genetic variants represents the driving force of biological evolution. Natural selection together with the available genetic variants determines the different directions that evolution takes. Isolation modulates the process of evolution. These interpretations are schematically represented in the upper part of Figure 1.

Thanks to microbial genetics and genomics it has become possible to investigate the molecular processes that are sources of genetic variation. On the one hand, bioinformatic comparison of nucleotide sequences from more or less closely related organisms can reveal the accumulated alterations within the genomes since their evolutionary separation, as well as, at lower levels of genome organization, within a functional domain, a single specific gene or a group of functionally related genes. This can allow the researcher to postulate the molecular mechanisms that are likely responsible for having generated the observed sequence alterations and the observed phenotypic changes. On the other hand, it is also possible to experimentally document single steps of nucleotide sequence alterations, particularly with small, microbial genomes. The results of such investigations are summarily represented in the lower part of Figure 1. Obviously, a relatively large number of specific molecular mechanisms are at the source of the overall genetic variation (Arber, 1997). These insights allow us to formulate a theory of molecular evolution (Arber, 2003, 2007).

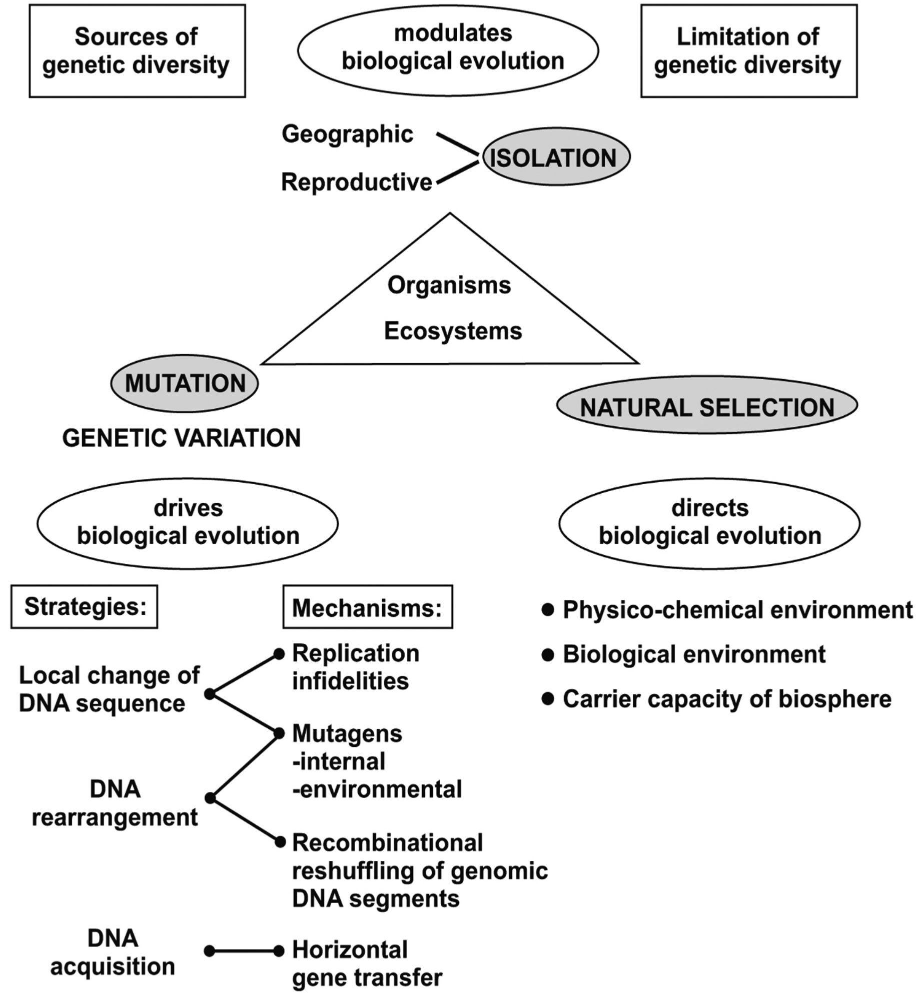

Sources of genetic diversity
modulates biological evolution
Limitation of genetic diversity
Geographic
Reproductive
ISOLATION
Organisms
Ecosystems
MUTATION
GENETIC VARIATION
NATURAL SELECTION
drives biological evolution
directs biological evolution
Strategies:
Mechanisms:
Local change of DNA sequence
DNA rearrangement
DNA acquisition
Replication infidelities
Mutagens
-internal
-environmental
Recombinational reshuffling of genomic DNA segments
Horizontal gene transfer
● Physico-chemical environment
● Biological environment
● Carrier capacity of biosphere

Figure 1. Schematic representation of elements involved in biological evolution and of the mechanisms and natural strategies of genetic variation.

3. Natural Reality Actively Takes Care of Biological Evolution

The specific molecular mechanisms of generation of genetic variants identified so far, some of which will be discussed in more detail below, can be classified into three natural strategies of genetic variation. Each of these strategies contributes qualitatively differently to the process of biological evolution.

3.1. *Local Sequence Change*

As it is shown in the lower left part of Figure 1, important contributions to the strategy of local change in the DNA sequence come from various types of replication infidelities that can, for example, result from tautomeric forms of nucleotides, from replication slippage or from a certain degree of chemical instability of nucleotides. Another source of local sequence change can be an impact of a chemical mutagen. It is well known that these kinds of mutational events would seriously threaten genetic stability if there were not could any enzymatic repair systems that could detect latent states of mutagenesis and reestablish the parental nucleotide sequence. Local sequence changes can occasionally give rise to the improvement of an available biological function or very rarely also to a new biological activity.

3.2. *Rearrangement of DNA Segments*

A second natural strategy of genetic variation is the intragenomic rearrangement of segments of DNA. In general, recombination enzymes catalyze the rearrangements that can result in the duplication, deletion, inversion or translocation of a DNA segment, or else in other kinds of reassortments. Occasionally, novel combinations can lead to a gene fusion or to the fusion of an open reading frame with an alternative sequence for the regulation of gene expression.

3.3. *DNA Acquisition*

The third natural strategy of genetic variation is DNA acquisition in which the genome takes up a segment of foreign DNA that had invaded the cell in question by horizontal/lateral gene transfer. This process is, in fact, at the basis of experimental microbial genetics and it is therefore well understood. More recently, horizontal gene transfer has also been identified to occur in higher organisms. The quality of DNA acquisition for biological evolution can be seen in a sharing by the recipient organism in successful developments made by others. Various, mostly enzymatically mediated limitations of DNA acquisition keep the frequencies of DNA acquisition low, thus insuring sufficient genetic stability.

3.4. *Involvement of Products of Evolution Genes and of Non-Genetic Elements in the Generation of Genetic Variants*

The theory of molecular evolution postulates that in living organisms all three natural strategies of genetic variation contribute with their differing qualities to the steady process of biological evolution at the population level. Products of specific genes, which are called evolution genes, are involved in most mutagenesis events of all three strategies. However, non-genetic elements are also involved, such as a certain degree of chemical instability of nucleotides and an intrinsic structural flexibility of biologically active molecules. Other non-genetic elements with influence on mutagenesis are environmental mutagens and random encounter. In summary then, intrinsic properties of matter together with the products of evolution genes cause spontaneous genetic variations at fine-tuned frequencies. This insures both a relatively good genetic stability of the individual organisms and enough genetic variants in populations to allow for a steady evolutionary progress of the population.

The products of some of the evolution genes such as repair enzymes and restriction enzymes modulate the frequencies of genetic variation to tolerable rates. Other evolution gene activities act as variation generators; they do this also at tolerable rates. It is postulated that these two kinds of evolution gene activities have become fine-tuned in their own past evolution by second-order selection exerted at the population level (Weber, 1996).

3.5. *The Nearly-Stochastic Nature of Genetic Variation*

This leads us to raise the question whether genetic variation is brought about by directed processes. In other words, does a bacterium, for example, possess a sensory organ to identify what kind of genetic change in which gene could allow the cell to adapt to an encountered altered environment. As far as we know, generally this scenario does not correspond to reality. Rather, spontaneous genetic variation occurs largely at random and not directively. As a matter of fact, only a relatively small minority of genetic variants turn out as favorable in comparison with their parental form. Many more spontaneous genetic variants provide to the concerned organism a selective disadvantage, and many other sequence alterations turn out to be neutral and silent under selective pressure.

4. Selected Examples of Molecular Mechanisms Generating Alterations in DNA

4.1. *Nucleotide Substitution Caused by Tautomeric Forms of Nucleotides*

Tautomery of nucleotides is known as one of the sources for nucleotide substitution, which is a local sequence change. In its standard form, adenine (A) pairs with thymine (T) in the double-stranded DNA molecules. Upon replication of DNA, the two daughter molecules will carry A-T pairs at the same sites as their parent, as long as the involved nucleotides are in their standard form. Occasionally, adenine can assume for a short moment its tautomeric imino form. Under these conditions, it cannot form a pair with thymine but it can do so with cytosine. In this latter case, as soon as adenine shifts back to its standard form, a mispairing becomes obvious. As we have already discussed, such mispairings can normally be rapidly repaired by appropriate enzymes. But occasionally, the mutation may become fixed as a nucleotide substitution that will later be transmitted to the progeny DNA molecules. In the scientific literature, this is often called a replication error. For conceptual reasons, I disagree with such interpretation. To my mind, natural reality rather makes use of the structural flexibility of tautomeric forms for the occasional stochastic production of nucleotide substitutions. The tolerable frequency of such local mutagenesis is thereby controlled by the efficiency of the enzymatic repair systems. In this case, the stochastic nature of mutagenesis is brought about by the non-genetic property of structural flexibility of the nucleotide, while the genetically encoded repair system modulates the frequency of such mutagenesis.

4.2. *Enzymatic Variation Generators*

Let us now direct our attention to two well-documented examples of enzymatic variation generators involved in the occasional rearrangement of intragenomic DNA segments.

4.2.1. *Transposition of Mobile Genetic Elements*

Mobile genetic elements are segments of DNA that carry genetic information which occasionally promotes either the translocation of the element or other kinds of DNA rearrangements (Shapiro, 1983). One widespread class of mobile genetic elements in bacteria is called IS (for inserted sequences) element. The transposition of an IS element into an alternative location in a genome can sometimes inactivate an important genomic function; in other cases, the process may have less drastic or even no observable effects. Quite rarely, as in most other mutagenesis events, the transposition activity may have a beneficial consequence.

Fig. 2 shows the result of an experiment in which lethal mutations in the genome of the bacterial virus P1 were screened for. The P1 genome can reside for prolonged times in the bacterial host as a provirus without producing viral particles. Only occasionally virus reproduction becomes spontaneously induced in one of the lysogenic cells. However, by experimental intervention, for example by UV-irradiation, one can induce virus reproduction in almost the entire bacterial population. A screening for bacterial subpopulations that do not produce viruses upon induction can reveal the presence of a lethal mutation in the resident viral genome. For the experiment reported in Figure 2, a population of bacteria that carried the P1 genome was propagated for several months under normal growth conditions and with periodical dilution into fresh medium (Sengstag and Arber, 1983). About 1% of the cells could then be shown by an appropriate screening not to produce active P1 virus any more. These cells still carried the P1 genome and about 95% of these independent isolates had suffered in the P1 genome an insertion of one of the IS elements that reside in the bacterial genome (most often it was the IS2 element). Obviously, in all these cases the IS insertion must have inactivated a viral function essential for viral reproduction. The P1 virus had suffered a lethal mutation. The upper part of Figure 2 shows the crude location of independent IS insertions within the 90,000 base pairs long P1 genome that is known to be densely populated with essential viral genes. Clearly, not all regions were used as insertion sites with comparable frequencies. Individual insertion sites of bacterial IS elements within the hottest region of IS insertion are shown in the lower part of Figure 2. Each of nine sequenced IS2 insertions had occurred at another location and these locations did not revel sequence homologies.

In a later experiment, either a stretch of the very hot region or a stretch of a cold region for IS2 insertion of the P1 genome was cloned in a plasmid

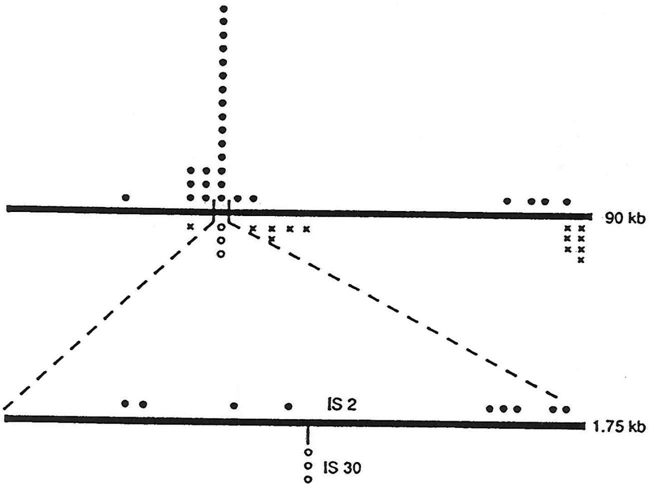


Figure 2. Location of independent IS insertions into the genome of bacteriophage P1 and resulting in mutants affected in the vegetative reproduction of the phage. The circular 90-kb genome of P1 is shown linearized in the upper part. Dots shown above the genome identify independent IS*2* insertions, crosses shown below the genome refer to insertions of IS*1*, IS*3*, IS*5*, Tn*1000* and circles to IS*30* insertions. As shown in the lower part, each of nine sequenced IS*2* insertions into the hot region for IS*2* transposition had occurred into a different sequence, while the three independent IS*30* insertions had occurred between the same base pairs and both orientations had been used. After Sengstag and Arber (1983).

gene vector (Sengstag and Arber, 1987). The resulting hybrid plasmids were then introduced into bacterial cells. Upon subsequent culture of these cells the hot region was still hot for IS2 insertion, while the cold region from the P1 genome was still cold at its plasmid location. In conclusion, there must be some still unknown regional characteristics in DNA molecules that can render this region attractive for IS2 insertion. Interestingly, however, each of the resulting transposition derivatives suffers the IS2 insertion at an individual location; the insertion event appears to be stochastic.

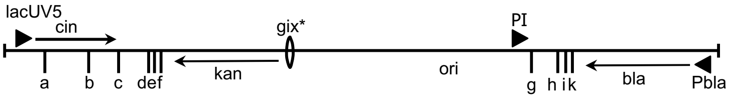
In contrast to IS2, the IS30 element shows a high degree of sequence specificity for its transposition. In the experiment reported in Figure 2, IS30 was found inserted into the P1 genome in three independent derivatives. All three insertions had occurred precisely at the same site of the P1 genome, twice in one direction and once in the opposite direction. IS30 is a good example for a mobile genetic element inserting with high probability into a specific sequence of nucleotides. With much lower frequencies, however, IS30 can also insert into other sites of DNA molecules.

With regard to its evolutionary relevance, IS transposition into many different genomic sites is of course more important than a nearly reproducible insertion into a specific and preferred DNA sequence.

4.2.2. *Site-Specific DNA Inversion Can Also Involve Secondary Sites of Inversion*

As the used terminology indicates, in site-specific DNA inversion a given segment of DNA can become inverted at its given location upon an enzymatic interaction (Glasgow *et al.*, 1989). The borders of this DNA segment are nearly-specific 'consensus' nucleotide sequences. In the well studied case of a DNA segment carried in the genome of the bacterial virus P1 the consensus sequence is 26 base pairs long and has a directedness in spite of a partial dyad symmetry. Two such consensus sequences carried in opposite directions flank a DNA segment that carries two different partial genes for determining the host range of the P1 virus. Just outside of one of the borders of the concerned DNA segment is another part of genetic information for the host range of the virus. In a so-called flip-flop process the DNA segment located between the two sites for inversion becomes periodically inverted. In both of the two possible orientations of the invertible DNA segment the host range gene product is active, but the host range of the virus differs for the two possible states of inversion (Iida, 1984). This is explained by the fact that a partial reading frame for the host range gene is constant and located outside of the invertible segment, while two alternative other parts of the reading frame are carried in opposite directions at the two ends of the invertible DNA segment. DNA inversion brings about an alternative fusion of the partial reading frames. Each of the two resulting fused genes determines its own specific host range of the P1 virus.

An interesting evolutionary contribution of such flip-flop systems of DNA inversion is brought about by the fact that, with much lower frequencies than the flip-flop reaction, the enzymatically mediated DNA inversion can also occur at so-called secondary sites of inversion. Figure 3 reports the results of an experiment in which an appropriate plasmid had been constructed with just one consensus inversion site (here called gix*). Proximate to the gix* site the plasmid carried an open reading frame for kanamycin resistance (kan), but without an expression promoter. As a matter of fact, the plasmid carried two promoters for gene expression (lacUV5 and PI) in the



	gix*	TTCctgTAAACC	AA	GGTTTTgGATAA	
a)	gix*Q391	Tatcacctctgg	AA	aGcaTacGgagt	(1)
b)	gix*Q390	aTagatgAttgC	AA	ccTgcTtacggg	(1)
c)	gix*Q389	cgCctgatAcag	AA	cGaaTTgctTgc	(2)
d)	gix*Q387	TTgggcTtcgga	At	cGTTTTccgggA	(5)
e)	gix*Q386	TTCTatgAAAgg	tt	GGgcTTcGgaAt	(1)
f)	gix*Q385	ccgcCtTctAtg	AA	aGgTTggGcTtc	(2)
g)	gix*Q384	actcttTttcCg	AA	GGTaactGgctt	(2)
h)	gix*Q388B	aTCTaggtgAag	At	ccTTTTtGATAA	(1)
i)	gix*Q393	aaaTtaaAAAtg	AA	GtTTTaaatcAA	(6)
k)	gix*Q392	aagTttTAAAtC	AA	tcTaaagtATAt	(1)

Figure 3. Nucleotide sequences used as secondary crossing over sites in Cin-mediated site-specific DNA inversion. On plasmid pSHI383 rare DNA inversion between the natural crossing over site *gix** and a secondary crossing over site brought the expression of the kanamycin resistance gene *kan* under the control of either promoter lacUV5 (sites a to f) or promoter PI (sites g and h). The plasmid also underwent unequal cointegration using *gix** and either site i or k and resulting in the fusion of *kan* with the operon under control of promoter P*bla*. Nucleotides corresponding to the *dix* consensus sequence of efficient crossing over sites are shown as underlined capital letters. Numbers in parenthesis refer to the number of independent isolates having used the crossing over site in question. The data were pooled from Iida and Hiestand-Nauer (1987), and the figure is reproduced from Arber (1995).

direction opposite to the kan reading frame, while a third promoter (Pbla) could not promote kan expression in view of its distant location. After a prolonged propagation of this plasmid in bacteria, kanamycin resistant derivatives were selected and the structures of their plasmids studied. Many derivatives had undergone DNA inversion or unequal cointegration between sister plasmids. In all of these cases analysed, recombination had occurred between the gix* consensus site and a secondary site (also called quasi site) on the plasmid (Iida and Hiestand-Nauer, 1987). Figure 3 shows the locations of ten different observed secondary sites and their nucleotide sequences (Arber, 1995). Capital letters identify nucleotides that still correspond to the consensus sequence. Considerable deviations from the consensus sequence are obvious and they vary from case to case. In this experiment a total of 22 independent kanamycine resistant insertion derivatives carrying one of the promoters for gene expression in front of the kan reading frame were sequenced. Although all studied invertants were independent, the secondary site (i) was used six times and site (d) five times. Three other sites (c, f and g) were used twice, while the remaining five identified secondary sites were used only once. In conclusion, many DNA sequences can serve as secondary sites for DNA inversion and plasmid cointegration with some statistical reproducibility: DNA inversion can occur repeatedly at a given secondary site, although always with very low frequency.

As it is the case for the much more frequent flip-flop reaction described above, the use of secondary sites of inversion depends on the presence of the specific DNA invertase enzymes (in this case the product of gene cin) and of a system for recombinational enhancement. Therefore, in all cases the observed DNA inversions are mediated by enzyme interactions.

In subsequent experiments it could be shown that DNA inversion involving secondary sites of inversion and leading to novel sequence fusions occurs also within the bacterial genome (Rozsa *et al.*, 1995).

What could then be the evolutionary role of DNA inversion systems? In the discussed experimental set-up this cannot be the relatively efficient flip-flop reactions which alternatively activate one or the other of the two observed viral host ranges. The two possible host ranges provide to the viral population a widening of its host range, but the flip-flop process does not really represent an evolutionary progress. In contrast, the much rarer use of secondary sites for the DNA inversion can provide a large number of possibilities for novel gene fusions as well as of fusions of open reading frames with alternative promoters of gene expression. Occasionally, one such novel fusion may provide to the concerned organism a selective advantage, thus contributing to the evolutionary progress. Similar to DNA deletion, DNA inversion thus represents a natural way to fuse previously disconnected functional domains, and some of the resulting recombinants may, by chance, carry out a novel beneficial function. Compared with DNA deletion, DNA inversion has thereby the advantage that no DNA sequences are lost from the genome. For these reasons, I consider enzymatically mediated DNA inversion as an effective natural contribution to the evolutionary progress. The involved enzymes act actively, but at low frequencies, as variation generators. I consider the genes that promote these reactions as evolution genes.

4.3. *Several Fine-Tuned Systems of Horizontal Gene Transfer Contribute to Microbial Evolution*

Microbial genetics originated in the 1940s when several basic mechanisms for horizontal (also called lateral) gene transfer became known. These are the transformation with free DNA molecules (Avery *et al.*, 1944), conjugation mediated by fertility plasmids (Lederberg, 1947) and virus-mediated transduction of bacterial DNA segments (Zinder and Lederberg, 1952). In the meantime, we know that products of specific genes are involved in all of these processes. Some of these genes can be classified as variation generators, others as modulators of the frequencies of genetic variation. For example, restriction-modification systems render horizontal gene transfer permissive within populations of bacteria belonging to the same restriction-modification system, while they largely reduce, but not completely suppress, the success of acquisition of DNA from more unrelated microorganisms (Arber, 1965, Arber and Linn, 1969). The stochastic nature of DNA acquisition can be seen at several levels of the process: in the mobilization of donor DNA for horizontal gene transfer, in the random encounter of the transferred DNA with a potential receptor cell and in the chance of the transferred foreign DNA to become part of the receptor genome.

Referring to section 4.2.1, it should be mentioned that bacterial IS elements do not only play an evolutionary role in intragenomic DNA rearrangements, they also contribute to horizontal gene transfer by their recombinational activities for the mobilization of chromosomal genes, as well as for the insertion of transferred genes into the receptor genome (Iida *et al.*, 1983). This had first been documented for genes providing antibiotic resistance and it has later been shown to contribute also to the horizontal transfer of other genes.

5. Conclusions and Outlook

Biological evolution is a steady, long-term process. It is driven by spontaneous genetic variation and this is largely, although not entirely, stochastic. Many specific molecular mechanisms contribute with their specific ways and qualities to the generation of occasional genetic variants. Some of the driving forces of genetic variation are exerted by products of evolution genes, while other genetic variations can be assigned to effects caused by intrinsic natural properties of the non-living world. Overall, genetic variation is thus understood to result from a tight cooperation between non-genetic elements, on the one hand, and products of specific evolution genes, on the other hand. Evolution genes also keep the frequencies of genetic variation low; this insures a relatively high genetic stability in populations. It is assumed that in their long past history, the evolution genes of today's organisms have become fine-tuned for their evolutionary tasks by second-order selection at the level of populations.

5.1. *How Predictable is the Evolutionary Progress?*

A long-term prediction that one can make is that biological evolution will continue to go on as long as the organisms can find environmental niches in which life can persist. Even if some biodiversity will be lost, the intrinsic power for evolutionary progress will steadily replenish biodiversity as it has done to create the present degree of biodiversity. However, expansion of biodiversity becomes limited by the carrier capacity of the planet Earth.

Almost nothing else can be accurately predicted for biological evolution. On the basis of today's knowledge one can confess not to be able to make precise predictions on naturally upcoming events of genetic variation; in the best of cases only statistical predictions can be made. Changes in environmental conditions affecting natural selection cannot be predicted accurately either, nor is this possible with regard to isolation phenomena. Biological evolution is an open system.

In this context, it can be considered as good news that we can predict that all encountered living organisms must possess a certain degree of genetic stability. If this were not insured, organisms with unstable genomes might have already disappeared from our planet by natural selection.

In turning our attention to those evolution genes that act via their enzymatic products as variation generators, one can realize, with surprise, that in contrast to many other genes, such as housekeeping genes, variation generators neither act reproducibly from case to case, nor efficiently (Arber, 2005). Otherwise, they would not be effective generators of genetic variations. This is illustrated by the two described examples of bacterial IS elements and of site-specific DNA inversion systems using secondary sites of nucleotide sequences for inversion. The underlying gene products are bona fide enzymes that exert their evolutionary activities with very low frequencies and at a large number of different sites in the genome.

5.2. *World View Aspects and the Duality of the Genome*

In the past, many textbooks on genetics and biological evolution described genetic variation as due to errors, mistakes and illegitimate processes. In the light of an updated theory of molecular evolution this interpretation has to be replaced by a more pro-active attitude in the interpretation of available data. According to the views that I have presented in this article, biological evolution is seen as an active process, in which natural reality cares by a kind of self-organization for the steady preparation of novel genetic variants that are submitted to natural selection. Since the living conditions vary both spatially and temporally, a stochastic provision of genetic variants can make sense, since it can allow for beneficial adaptation at the population level to a number of different encountered environmental conditions.

One of the philosophical consequences of the postulate of evolution genes is the duality of the genome (Arber, 2005). We can realize that many, but not all, of the genes carried in a genome serve to satisfy the needs of the individual life of the carrier of the genome. They contribute to the fulfillment of the individual life. In contrast, the evolution genes contribute, at the population level, to a steady expansion of life, to the adaptation to novel living conditions and to a steady replenishment of biodiversity. This is welcome news of hope for us human beings, although we cannot precisely predict how the numerous forms of life on our planet will further develop in the long term.

REFERENCES

Arber, W. (1965), Host-controlled modification of bacteriophage, *Annu. Rev. Microbiol.*, 19, 365-378.

Arber, W. (1995), Genetic bases for evolutionary development of microorganisms, In: van der Zeijst, W.P.M. *et al.* (eds.), *Ecology of Pathogenic Bacteria; Molecular and Evolutionary Aspects*, Royal Netherlands Academy of Arts and Sciences, pp. 3-13.

Arber, W. (1997), The influence of genetic and environmental factors on biological evolution, *Commentarii*, Pontifical Academy of Sciences, vol. IV, 81-100.

Arber, W. (2003), Elements for a theory of molecular evolution, *Gene*, 317, 3-11.

Arber, W. (2005), Dual nature of the genome: Genes for the individual life and genes for the evolutionary progress of the population. *IUBMB Life*, 57, 263-266.

Arber, W. (2007), Genetic variation and molecular evolution, In: Meyers, R.A. (ed.), *Genomics and Genetics*, Wiley-VCH, Weinheim, vol. 1, 385-406.

Arber, W. and Linn, S. (1969), DNA modification and restriction, *Annu. Rev. Biochem.*, 38, 467-500.

Avery, O.T., MacLeod, C.M. and McCarty, M. (1944), Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *J. Exp. Med.*, 79, 137-158.

Glasgow, A.C., Hughes, K.T. and Simon, M.I. (1989), Bacterial DNA inversion systems, In: Berg, D.E. and Howe, M.M. (eds.), *Mobile DNA*, American Society for Microbiology, Washington, DC, pp. 637-659.

Iida, S. (1984), Bacteriophage P1 carries two related sets of genes determining its host range in the invertible C segment of its genome, *Virology*, 134, 421-434.

Iida, S. and Hiestand-Nauer, R. (1987), Role of the central dinucleotide at the crossover sites for the selection of quasi sites in DNA inversion mediated by the site-specific Cin recombinase of phage P1, *Mol. Gen. Genet.*, 208, 464-468.

Iida, S., Meyer, J. and Arber, W. (1983), Prokaryotic IS elements, In: Shapiro, J.A. (ed.), *Mobile Genetic Elements*, Academic Press, Inc., New York, pp. 159-221.

Lederberg, J. (1947), Gene recombination and linked segregation in *E. coli*, *Genetics*, 32, 505-525.

Mayr, E. (1982), *The growth of biological thought: Diversity, evolution and inheritance*, Harvard University Press, Cambridge MA

Rozsa, F.W., Viollier, P., Fussenegger, M., Hiestand-Nauer, R. and Arber, W. (1995), Cin-mediated recombination at secondary crossover sites on the *Escherichia coli* chromosome, *J. Bacteriol.*, 177, 1159-1168.

Sengstag, C. and Arber, W. (1983), IS2 insertion is a major cause of spontaneous mutagenesis of the bacteriophage P1: non-random distribution of target sites, *The EMBO J.*, 2, 67-71.

Sengstag, C. and Arber, W. (1987), A cloned DNA fragment from bacteriophage P1 enhances IS2 insertion, *Mol. Gen. Genet.*, 206, 344-351.

Shapiro, J.A. (ed.) (1983), *Mobile Genetic Elements*, Academic Press, Inc., New York.

Watson, J.D. and Crick, F.H.C. (1953), Molecular structure of nucleic acids. A structure for deoxyribose nucleic acid, *Nature*, 171, 737-738.

Weber, M. (1996), Evolutionary plasticity in prokaryotes: a panglossian view, *Biol. Philos.*, 11, 67-88.

Zinder, N. and Lederberg, J. (1952), Genetic exchange in Salmonella, *J. Bacteriol.*, 64, 679-699.

Session IV

PHILOSOPHICAL AND SOCIETAL ASPECTS

EPISTEMOLOGICAL STUDY OF THE VOCABULARY OF PREDICTION IN SCIENCE AND THEOLOGY

JEAN-MICHEL MALDAMÉ

Since this symposium is dedicated to the study of predictability, my paper will endeavour to show how the terms used in scientific discussions, in the philosophy of nature and in theology keep interacting. This will go to prove that it is important to be able to mark distinctly the meaning of terms according to the levels of discussion and conceptualisation and thus avoid the misunderstandings which today seem to have caused the dialogue between science and faith to reach a dead end.

1. *The Return of an Old Quarrel*

In Western thought, a clear distinction between scientific knowledge and theological considerations has been the universal rule. It was quite evident that scientific work should not be mingled with religious or ideological considerations. It was a manner of respecting the objectivity of scientific knowledge, built upon the exacting demands of the experimental method – where the subjectivity of the observer must not interfere with the results. Such an attitude was an essential element of the freedom of research.

Now, this situation has drastically changed today through the influence of religious groups which play an important part in North-American society.[1] Their project of reforming society includes religious elements, which is logical, since religious convictions and practices are integral parts of human

[1] The question is complex and one must carefully distinguish between the various trends of ideas or opinions which introduce confusions between religious and scientific discourses. As far as I am concerned, I could not possibly mistake creationism for fundamentalism; or place those in favour of the Intelligent Design in the category of creationists: it would be a way of evading discussion.

identity. What is new, is that in the name of religion, the supporters of *Intelligent Design* claim that they are able to settle questions discussed by scientists: by rejecting the theory of evolution, some of them even claim that they can judge the value of a scientific theory by referring it to biblical texts.

Most scientists have said that such an interpretation has revived the old conflicts between science and faith, since a religious conviction has imposed on scientists what they must think. Since one the elements of the present debate concerns the relationship between chance and finality, it seems to me that this convention concerning scientific predictability is a good opportunity to examine in a balanced way the relationship between science and monotheistic theology: as a matter of fact, one should avoid the simplistic approaches of 'concordism'[2] or 'discordism',[3] when one tries to examine how scientific discourse and theological discourse get on together – and thus denounce the confusions which are caused by those who support *Intelligent Design*, when they oppose divine design and chance in the name of finality.

2. *Causality and Scientific Prevision*

Predictability has always been one of the major concerns of mankind – in order to know in advance about the return of seasons, to evaluate climactic variations or to assess what resources were necessary for survival. Such a requirement was revived at the birth of rational thought in a vision which grants primacy to reason in human actions. This primacy of reason has kept away from the plurality of divine beings that haunt ancient mythology. In order to establish social and political order as well as learning, the Greeks introduced the concepts of law, order and reason – placed under the exacting demand of logical coherence.

Within this framework, Greek thinkers introduced the notion of cause.[4] Hence the need to explain turns facts and events into serial sequences

[2] A posture which claims that the inspired text and the results of science coincide word for word. The Catholic church has challenged this position, which was prevalent at the beginning of the twentieth century.

[3] It is the opposite attitude, which considers that science and faith don't have to mind or worry about each other – even when creation or providence are at stake.

[4] In Greek thought, one talks of science (épistèmè) when the cause (aitia) is known. The concept of cause refers to the fact that a 'being' (in the widest possible sense, meaning an individual, an event, a connection of any sort…) is dependent on another being. The notion of cause refers to the questions 'why?', 'how', 'what is it' or 'what is it made of?'.

which connect to one another with the passing of time. They develop in a mode which is made clear through the tenses of verbs: past, present and future. The present is the outcome of the past and the future is understood as the result of the present action. The link between facts is a causal link. When a cause is established, effects necessarily follow. It cannot be otherwise, as the saying goes: 'the same causes produce the same effects'.

This ideal of perfect knowledge has led scientific thought to introduce a fundamental distinction between predicton and speculation, between science and opinion.

When scientific knowledge is at work, it is possible to foresee what is coming, since the whole process is governed by necessity. For the Ancients, science proceeded through the knowledge of necessity. But there are situations which do not correspond to this ideal requirement. Facts are not necessary, they could be different, they are contingent and no demonstration can be made about them with any certainty. In this case, we are not talking of science, but of opinion: prediction is impossible, only conjecture remains.

This philosophy of knowledge has a logical aspect, since ideal knowledge rests on the conviction that there must be a relation between things and thought. Logical necessities are necessities of the human being. This entails a theory of demonstration.[5]

Since truth does not reside only in concepts, but in judgements, one must examine the status of propositions. A proposition is true when, if affirmative, it says what is, or if negative, what is not. Hence it follows that propositions concerning present or past things are necessarily either true or false, since they always refer in the present or in the past to something with which they are or are not in agreement. On the other hand, if propositions cannot refer to anything in reality which either confirms or denies them, they are neither true nor false: they are neutral. This neutrality characterises propositions which refer to what is called the 'contingent future'.

The expression applies to what is to come in the future, but does not belong to the field of necessity. For Aristotle, such facts cannot be foreseen, so that the propositions which concern them can neither be true nor false. They are neutral.

Thus, there are two categories of facts: the former concerns what exists according to the necessity of natural laws, or events which have taken place.

[5] See Aristotle, 'Seconds Analytiques', I,2,71b, tr. Jean Tricot, *Organon IV*, Vrin, 1938, pp. 7-8.

The latter concerns facts which occur without reference to necessity. It is possible to assert the truth or untruth of what has happened; but one cannot assert the truth or untruth of what is yet to happen outside the scope of necessity.[6] This does not create any dificulty, as long as one stays at a logical level, since one accepts that there are such things as natural assertions[7] and that there is a limit to human knowledge; but it is of some importance when such a conception of knowledge is made to apply to theology.

3. *Providence and Predestination: Contingent Futures*

The following discussion only concerns monotheistic theology. Acknowledging plural divine beings only means extending what's taking place on earth into heaven. But acknowledging a unique God – as do the philosophers and Abrahamic religions – invites one to reconsider the preceding distinctions in the light of the problems raised by the existence of chance. Chance has been defined as the meeting point of independent causal series.[8] The random character of events follows from their independent causes, which make prediction impossible.

In order to enter the theological debate, it may be appropriate to resort again to a classical example used by philosophy teachers: two slaves are sent by their master on an errand. They start independently from one another, each without knowing what the other has to do. Supposing they meet in the town square; they will say that their meeting was due to chance, and happened at random.From their point of view, their encounter was the result of two series of independent causes. Not so for the master, who – in so far as he has sent them to the same place at the same time, is entitled to believe that they were bound to meet. Their meeting does not have the same nature for

[6] The question has been discussed at length in the Treatise of Interpretation, and illustrated by the example of a naval battle supposed to take place in the future. Because it is not certain that the naval battle will take place at all, since it depends on a number of factors which include human freedom, it is impossible to pass a judgement on the proposition: 'the battle will take place tomorrow'. Since it is a contingent fact, one cannot with any certainty ascertain the truth or untruth of the proposition. Things go differently where a naval battle of the past is concerned: propositions concerning it can be classified into two categories of true and false propositions.

[7] This Aristotelian reserved attitude is rejected by the Stoics, who talk of modal propositions.

[8] Aristotle's definition has been taken up by modern science under the influence of Cournot.

the master and for his slaves. This difference in points of view between master and slaves has been taken up in theology. In the case of a monotheistic confession of faith, the situation of men – caught up as they are in temporality – is different from that of God, who is outside Time. So that the question of chance and predictability has entered Christian theology, which claims, in accordance with the requirements of monotheism, that God sees everything. Therefore, the distinction between past, present and future events is not a radical one: everything is present, for Him. If one takes into consideration only the knowledge which God has of past and present facts, such a theological proposition does not offer any major difficulty. But if one considers that this knowledge can be creative, a difficulty arises: since God knows everything and since nothing can be kept unknown from Him, does it mean that such an action invalidates the contingent character of what it creates?

The answer to such a question has occasioned important research in theology, on the question of what has been called ‘contingent futures’. A thorough discussion of this academic disputation[9] is here out of the question. One can however observe that in a monotheistic tradition, two conceptions of divine action are at work.

4. *The Action of God and His Design*

In order to answer the question asked by the supporters of the *Intelligent Design* theory, it is possible to consider two opposite traditions.[10]

According to the first tradition, the notion of omnipotence designates the absolute character of the power God, who is supposedly able to do all that he wants without being stopped by anything. This school of thought stresses the word ‘all’, which refers both to the universal character of what is, and to the universal character of what could be – and even to what lies outside the scope of human imagination. Nothing can limit the action of God, which evades all rational explanation.

[9] It is not possible to reproduce the debates here, but suffice it to say that the discussions have encouraged research concerning precise vocabulary and definition of concepts. The debate has been most sharp at the University of Louvain about what is known as ‘the quarrel of contingent futures’. See Leon Baudry, *La Querelle des futurs contingents (Louvain 1465-1475)*, Paris, Vrin, 1950. Throughout the debates, academics have distinguished what they call in Latin *de re*, in other words what refers to things, from *de dicto*, namely words, a current discussion in the debates concerning quantum physics.

[10] See the anthology by Olivier Boulnois, *La Puissance et son ombre, de Pierre Lombard à Luther*, Paris, Aubier, 1994.

According to the second tradition, the action of God cannot be dissociated from the notion of wisdom. This implies that the order of elements and the proportion between causes and effects are also taken into consideration. Thus, God's will is motivated by Good and regulated by the notions of order and coherence.

The first tradition, for which the power of God is absolute, does not take into consideration the demands of coherence between natural phenomena. The contingent quality of the facts of nature is an irreducible factor, because God's interventions are limitless. He is free from any restraint arising from reason.

The second tradition, on the other hand, makes the power of God subservient to His wisdom. Then things are proportionate to one another, and the links between causes and effects are well-ordered. Such a conception does away with the notion of God's arbitrary behaviour. It gives sense and meaning to the expression: 'divine design'. But again, the interpretation of such an expression calls for proper judgment.

In the first perspective, the divine design can be recognized only if it cancels the contingent nature of facts or natural events. The notion of divine design clashes with the recognition of chance. In the second perspective, by favouring wisdom and therefore reason, contingency is not invalidated by the recognition of God's action. In fact, phenomena occur according to their nature. In such a perspective, the notions of creation, divine design and providence do make sense. Chance and God's action are not antinomic, because the latter does not distort natural events and respects the rules of the possible. As far as I am concerned, I see no reason why one should oppose chance and God's design.

This respect for the nature of things by the Creator means that if an event is contingent, it does not stop being so because it is willed by Him whose action is at the very source of being.[11] It is willed as such, contin-

[11] The notion of contingency implies that all that god has created does not necessarily partake of the absolute or necessary nature of His being. Thus, in a well directed human action, he who acts adjusts his forces to what he is doing. He does not use all his potentialities to achieve what he is in the act of doing. Thus, an action does not always involve the same resources. Practical wisdom consists in adapting one's efforts to the work in progress. This remark could apply (by carefully respecting the differences between God and man) to divine action. When we describe it as supreme, we do not mean that the whole divine power is engaged to do away with contingency, and therefore with the random nature of phenomena. Here again, the notion of wisdom, which adapts causes to effects, must be favoured. Divine action respects the singularity of human beings and their links.

gent as it is. There is therefore no reason to oppose providence and the chance happenings of life, chance and God's design. Nor is there any reason to oppose the notion of creation to the synthetic theory of evolution. The action of God respects the laws of Nature. How could he do violence to the laws He himself has established? Creative action develops within the framework of time. It takes place through evolution. The general term of evolution has taken a particular meaning in science, when used to designate chance mutations.

The philosophy which we are expounding here discards any explanation referring to special interventions of God, which would distort the normal course of nature. It might be called the 'autonomy of creatures'. The theology which we are promoting rests on the conviction that God's action does not alter natural phenomena and leaves science totally free in its quest for explanations.

5. *Creative Action and the Autonomy of Creatures*

By admitting that God's action is not an intervention which alters the course of natural phenomena, we leave science in its own place. Of course, science does not know everything. But facing the unknown, and the enigma created by the emergence of life and the apparition of mankind, it is erroneous to appeal to an intervention of God which would alter the natural course of things. It is enough to realise that our knowledge is still limited. The theory of evolution accounts for what has been observed. It is open to new discoveries. It looks forward to them. It will renew itself thanks to discoveries to come, unforeseeable in the present state of our knowledge. But such a revision will not repudiate what is today established and verified; it will be a re-interpretation in a wider framework, through a more widely encompassing theory. The opposition between the synthetic theory of evolution and Christian faith is without foundation.

God's action respects the laws of Nature. God's action does not do violence to the rules which He himself has established. God's action then operates through the mechanisms revealed by the theory of evolution. Does creative action take place within the framework of time? It takes place through evolution. It is a general concept. It does away with an explanation which would allow special interventions by God, which would distort the normal course of nature.

At this stage, it is necessary to add a few refinements. First concerning creative action.

1. What is ordinarily meant by creation, is the very first moment of the temporal history of beings – normally represented by point zero on the standard cosmology timeline. It is in this perspective that the arguments of the supporters of Intelligent Design normally develop, on 'fine tuning' and the opposition between microevolution and macroevolution. According to them, God can only intervene without our knowing, or even counter to the natural course of things, in order to bring new solutions, orientate, maybe redress.

This narrow conception is not that of the Christian tradition, for which creation is an act of the present, always present. Providence is a quality of creative action, whose completion is inscribed within a temporal framework.

2. God's action must not be conceived as an intervention on subjects which is meant to orientate those subjects in a way which would not correspond to their being. It should be thought of as the gift of being to what is singular and makes up a whole – the world. It is a universe in the eyes of the scientist, in the sense that phenomena occur according to laws. It is called 'creation' from a theological point of view, when one realizes that it conveys one unique will. This approach lays stress on its unity and on the dynamic movement which drives it towards an accomplishment of some sort, without anything being distorted or the interaction of elements between themselves where chance has a part to play.

6. *The Question of Sense and Reality*

In this perspective, one can say that the question of sense is being raised within the process of evolution, without it being necessary to deny the random nature of singular occurrences by supposing that an intervention of God has somehow filled a gap of some sort.

Does the admission of contingency compel us to give up the possibility of speaking of a divine design? In order to do so, it is not enough to consider singular occurrences. One must consider the whole vital process described by the theory of evolution. It is a fresh way of looking at things. It takes the whole process into consideration and acknowledges its coherence. This is where the followers of *Intelligent Design* start from.

a) In the first place, those in favour of *Intelligent Design*, in taking note of the fact that the universe is in expansion,[12] address the question of sense,

[12] See Rodney D. Holder, *God, the Multiverses and Everything, Modern Cosmology and the Argument from Design*, Hampshire G.B. Ashgate, 2004.

passing from what deals with the direction of a movement to its meaning.[13] The anthropic cosmological principle extends this consideration and refines it, by considering that the universe is orientated towards the emergence of human consciousness.

b) Then, the supporters of the *Intelligent Design* develop the meaning of the word 'program'. They extract the term from the context of a mechanistic conception and make it apply to an intelligence at work at the beginning as well as at the end of the process.

c) If there is for them the possibility of an optimal solution starting from elementary conditions, it proves that there is an internal guidance which uses the openings of the possible for an optimal result, which goes beyond what scattered elements could produce at their own level.

This argumentation unfortunately rests on the opposition between chance and God's design. It ignores an essential point, concerning the nature of the theory of evolution. That theory makes it possible to retrace the history of life. It is voiced by those who are at the end of the process at work. They see links between the forces of nature and occurrences. They offer explanations which extend from the limited field of their studies to the whole phenomenon of life. It is important to admit that it is a retrospective vision. By placing oneself ideally at the beginning, it is not fair to say that one could foresee what happened. It is not pointless to foresee the future evolution of life, but those predictions are of random phenomena – the farther one goes from the present to look into the future, the less one can foresee what will come, as we well know if we think of the determinist chaos. Facing the unknown, it is important to admit that the existence of divine interventions is yet to be proved – and therefore one must avoid appealing to them.

Thus, in order to speak of a divine design in theology, we have shown that it is not necessary to oppose God and chance. The concept of chance is unfortunately too large. On must show that chance is not only a zone of 'unknowing', but that there are laws which make it possible to deal with it. We need to go through a last stage in our argument and consider philosophy immanent to mathematical probabilities.

[13] See John Polkinghorne, *Science and Creation*, London, SPCK, 1988; *Science and Providence*, London, SPCK, 1989; *Science and Christian Belief*, London, SPCK, 1994.

7. *The Notion of Probability*

The current reflexion of the more sensible supporters of *Intelligent Design* takes its roots in a reflexion on conditional probabilities.[14] It is by analysing the way in which an argument concerning conditional probability functions, that the supporters of *Intelligent Design* favour the existence of an intelligence at work in the world.[15]

Concerning this point, it is important to see that the reference of science to probabilities introduces a new perspective. The Ancients gave to the notion of Cause an ontological value. Modern science has abandoned this philosophy, considering that one should not reason in ontological terms. Before being used by sciences, the notion of probability was made clearer through debates concerning the relationship between science and opinion.[16] Whereas science proceeds in a demonstrative way and leads to certitude, the word probable qualifies opinion. The adherence to what is probable has been the object of interest of moralists, concerned as they were with actions corresponding to the standards of truth and rectitude, and of jurists, who had to take decisions in muddled situations. For want of turning to experience and demonstration, a probable opinion was a judgment approved by autority or by the testimony of respected judges. Probable opinion referred to plausibility.

The situation changed in the 17th century, when calculations were made to determine what was the most probable. Pascal's findings were resumed by Leibniz, who was the first to suggest the use of calculation in order to measure the degree to which a proof was valid.[17] At the end of the

[14] This theme has been dealt with at some length by Michael Behe, *Darwin's Black Box*, New York, Free Press, 1996. On this point, see the review by Eliott Sober, 'Intelligent Design and probability reasoning', *International Journal for Philosophy of Religion*, n. 52, 2002, pp. 65-80.

[15] See the book by one of the supporters of *Intelligent Design*, William A. Dembski, *The Design Inference. Eliminating Chance through small Probabilities*, Cambridge / New York: Cambridge University Press, 1998. On this question, see John Forster, *The divine Law-Maker, Lectures on Induction, Laws of Nature and the Existence of God*, Oxford, Clarendon Press, 2004.

[16] Ian Hacking, *The Emergence of Probability*, Cambridge University Press, 1975. Trad. Fr. *L'Emergence de la Probabilité*, Paris, Ed. du Seuil, 2002.

[17] This research shows how a new science benefits by the conceptual contributions of other fields; here, the science of action in the fields of morality, justice, and the management of goods.

17th century, Bernoulli synthetised those results and formulated a theory proposing a global vision of the notion of probability linked with a mathematical expression in the shape of a theorem.[18] Ever since, scientists have invented more powerful mathematical tools and, in so doing, have changed the nature of scientific work.

For the Ancients, at the school of Plato or Aristotle, the link between a cause and its effect is of an ontological nature, it is necessary. The mathematical expression is then a source of certitude. The consideration issuing from the development of a science founded on probabilities is quite different. By using no longer what is certain but what is probable as a support for its demonstration, science no longer says *what is*, but *what happens most often*. The questions asked by Hume express this change. Abstraction is replaced by induction. In the classical sense of the word, abstraction catches the essence of a phenomenon, and separates it from adventitious elements; in the modern sense of the word, induction is a generalisation which remains in the field of the probable. The words *law* and *cause* do not have the same meaning. In current science, the notion of predactibility has broken away from the deterministic vision whose ambition is to account for everything by necessary reason, without leaving anything outside of its scope.[19] A classical example of this is to be found in the preface by Laplace to his *Treatise on Probabilities*.

The notion of predictibility has radically changed. For that reason, the transfer made by the supporters of *Intelligent Design* from science to theology is not a rigorous step, because it does not respect the orders of knowledge: it unduly introduces theological developments within scientific discourse. It has been our concern to honour philosophy by talking of contingency and science by talking of random phenomena.

Conclusion

At the close of this rapid analysis, I would like to introduce a distinction between three ways of considering the action of God the Creator. One way seems to me quite traditional: the world has been created by God. The cre-

[18] On the question of probabilities, see the fundamental study by Jean Largeault, *Hasard, Probabilités, Inductions*, Toulouse, Université du Mirail, 1979.

[19] See Alexandre Kojève, *L'idée du déterminisme dans la physique classique et dans la physique moderne*, Paris, Librairie Générale française, 1990.

ating act does not consist in establishing a universe, which God would have then abandoned. It is a permanent presence.

A second approach strikes me as inadequate: I will describe it as apologetic, because it uses the inadequacies and uncertainties of science to propose an intervention of God. It is improper, because if and when we eventually understand, God becomes an 'unnecessary hypothesis'.

A third attitude seems to me unacceptable: it consists in finding arguments in the Bible to challenge or refute the scientific explanations which do not use word for word the biblibical text. Here is there a misunderstanding of the nature of the biblical text.

At the end of this paper, I would like to remark that these difficulties arise from a too narrow conception of the action of God. If you make God into an actor like other actors, you have to push God aside so that nature can act, and nature must be purely passive in order to obey God.

We have a different conception of God. The acknowledgment of his sanctity allows for the autonomy of his creatures and the play of nature's laws. Such a position is rooted in revelation. Revelation, far from limiting scientific activity, founds it by showing that God is greater than what religions and philosophies ordinarily admit. Revelation, far from giving ready-made answers to human research, underscores the importance of man's freedom, circumscribed as he is in a nature which has its own coherence and richness.

Those elements apply to nature; they could also be used to build a theology of human freedom. The more we acknowledge the greatness of grace, the better we understand that nature is at work according to its capacities in the adventure of salvation.

PREDICTABILITY, MEASUREMENTS AND COSMIC TIME

MICHAEL HELLER

1. *Introduction*

There exists the common consensus of both physicists and philosophers of science that empirical predictions belong to the core of scientific method. However, the claim that a piano encircles the planet Uranus along an elongated orbit, although empirically – in principle – falsifiable, never would be taken seriously. To define, from the methodological point of view, the nature of scientific predictions and their role in the sciences is not an easy task. For our purposes it is enough to emphasize that any truly scientific prediction (at least as far as physics is concerned) must follow from a theory, expressed in a mathematical form, and must refer to concrete measurement results. For a prediction to be a part of scientific method it is necessary to follow from a scientific theory. Even the most accurate predictions made by an oracle do not count in the sciences. We thus must have a theory *now*, and the prediction *directed to the future*. In this way, the directional flow of time seems to be involved in the very possibility of physics.[1]

And what about retrodictions? If we look at the history of physics, we can easily convince ourselves that retrodictions were as important as predictions. For instance, one of the classical empirical tests for the theory of general relativity, the perihelion motion of Mercury, was very well known to astronomers for a half of century, but this fact did not prevent regarding Einstein's 'prediction' as a major breakthrough in physics. The possibility of reconstructing a state of a physical system in the past (from the present theory) is as important, from the methodological point of

[1] I am treating physics as a typical example of empirical sciences.

view, as predicting it in the future. Only from the psychological point of view we are inclined to attach a greater value to predictions rather than to retrodictions.

It is interesting to notice that these simple remarks on physical methodology lead to a nontrivial cosmological conclusion: *physics, as a science, is possible only in a universe in which exists a local time.* It is clear that it is enough to have a *local* time, i.e., time defined in a neighborhood of the physicists making predictions or retrodictions. In such a time there exist two directions that may be arbitrarily labeled 'the past' and 'the future'. Strictly speaking, the time arrow, pointing to the exactly one direction as to the future, does not seem necessary.[2] Physical time is a time *measured* by a clock. Therefore, physics is possible only in the universe that admits the existence of a clock. One can hardly imagine a clock without it being localized at a certain place. Moreover, to have a *local* time means to have a space-time neighborhood in which a clock is situated. All clocks (and other measuring devices) to be usable by human physicists, must be macroscopic contraptions, or at least must have 'pointers' in the macroscopic world. Thus, the spatio-temporal structures we postulate are macroscopic structures. This, of course, does not exclude the possibility for the physicist to invent and develop theories regarding both the micro-world of atomic and subatomic dimensions, and the world on the cosmic scale, but all these theories have to be tested in our macro-world. As far as the possibility of doing physics is concerned, the macroscopic physics is essential.

The aim of the present paper is to make the above intuitions more precise and to look for some of their philosophical consequences.

2. *Space and Time Measurements*

Among physical measurements especially important are time and space measurements. Although time and space (length) units can be constructed from other physical quantities, space and time are usually regarded as belonging to the most 'primitive' physical magnitudes. According to all theories of macroscopic physics, physical processes unfold on a space-time

[2] This idea was elaborated in my paper: 'The Origins of Time', in: *The Study of Time, IV,* ed. by J.T. Fraser, N, Lawrence and D. Park, Springer, New York, 1981, pp. 90-93, and in the Lecture 4 of the book: *Questions to the Universe. Ten Lectures on the Foundations of Physics and Cosmology,* Pachart, Tuscon, 1986. The present paper is partially based on these works.

arena which, from the mathematical point of view, is a *differential manifold* (or *manifold*, for short). However, the concept of differential manifold as such is too poor a concept to serve as a suitable arena for physics: in the manifold structure there are no conceptual tools that would enable time and space measurements. To acquire such tools, the manifold structure must be enriched by superimposing on it another structure, called *metric structure*. There exist many metric structures and it is up to experiment to decide which of them is correct to model the real world. The present physical paradigm says that it is the *Lorentz* metric structure. More precisely, a four dimensional differential manifold, equipped with the Lorentz metric structure is the mathematical model for physical space-time. Within this model space and time measurements become meaningful operations.

Although in this paper we are essentially interested in the macroscopic space-time model, it is worthwhile to notice that it has been experimentally verified with enormous precision in the realm of microphysics. Predictions of the standard model of elementary particles presupposing this model have been verified with amazing accuracy at length scales of about 10^{-16} cm,[3] and the latest clocks, using a single ion, measure time with the (anticipated) precision of 10^{-18}.[4] We can expect that only below this threshold our manifold model of space-time breaks down. In fact, many works aiming at creating the fundamental theory of physics, predict that something like that should happen.

In the *Introduction* we have said that the minimal cosmological condition for making predictions is the existence of a local time. Geometrically, it is a very tolerant condition. As it is very well known, on every differential manifold a Lorentz metric *locally* always exists. But physics is more than geometry. As we shall see below, the construction of a physical clock requires an interaction of 'nonlocal' regions of space with each other. Moreover, physics regarded as the collection of physical laws, operates not only locally, in our corner of the Universe, but everywhere, i.e., globally. In any case, we should adopt a modern, more universalistic perspective, and at least consider the possibility of the existence of 'other physicists' somewhere in the Cosmos. Since physics can happen on a space-time manifold only if it is a Lorentz manifold, we should look for the necessary and sufficient conditions of the *global* existence of the Lorentz structure.

[3] http://ltp.web.psi.ch

[4] J.C. Bergquist, S.R. Jefferts and D.J. Wineland, 'Time Measurement at the Millenium', *Physics Today,* 54, no 3, 2001, 37.

And in this case the answer is well known. A Lorentz metric g exists on a differential manifold M if and only if a (smooth) non-vanishing direction field is defined on M. In other words, at each point of M there should be the possibility to distinguish two directions which we can arbitrarily label 'backward' and 'forward'. If we determine which is 'backward' and which is 'forward' (and if we do that is a smooth way) then the direction field becomes the vector field (to each direction we attach an arrow pointing to the 'forward', say).

The proof of the above theorem is by construction.[5] Let us suppose that on a manifold M there exists a Lorentz metric. With the help of this metric, we construct a light cone at every point p of M. In the interior of every such light cone we choose a vector (which is of course a timelike vector in this metric). This can be done is a smooth way. We thus obtain a smooth vector field on the manifold M, and the vector field obviously determines the direction field. And now let us suppose that on M there exists a nowhere vanishing direction field. There exists a simple recipe how to construct a Lorentz metric out of this direction field. This can be done in such a way that the direction field becomes timelike in this metric.

This result is striking. It can be interpreted by saying that a differential manifold is a suitable arena for physics if and only if every observer in it can distinguish two time directions. Perhaps the term 'observer' is here an exaggeration. Such an observer need not to have a Ph.D. in physics; it is enough for him/her to be equipped with a suitable feeling of time. In this sense, time is a precondition for physics.

This result should be understood correctly. The above theorem asserts that on a manifold M there exists a Lorentz metric globally if and only if on M there exists a nowhere vanishing direction field that can be interpreted as a sort of time feeling by every observer. The Lorentz metric exists globally, but it is enough for the required time to be local. Clocks carried by all local observers need not be synchronized. The only requirement is that two time directions change smoothly from one observer to another (so as the non-vanishing direction field be smooth). Of course, we can postulate the existence of a global time on the manifold M, but this requires an additional condition imposed on it. Surprisingly enough, this condition is also related to the possibility of doing physics on M.

[5] For details see: R. Geroch and G.T. Horowitz, 'Global Structure of Spacetimes', in: *General Relativity. An Einstein Centenary Survey,* ed. by S.W. Hawking, W. Israel, Cambridge University Press, Cambridge, 1979, pp. 212-293, especially pp. 218-220.

3. *Global Time and Stable Measurements*

Everyone who ever had to do something with performing physical measurements knows that they give results only within certain limits of accuracy (even without taking into account quantum indeterminacies). And such measurements give us valuable information about the world on which all of our science is based. This very fact also contains an information about the structure of the Universe. Imagine a 'malicious universe' in which even the slightest change in the measurement result leads to drastically different physical theories. In such a universe no theory could ever be empirically verified or falsified. Small errors inherent in every measurement would spoil the effectiveness of the empirical method. Since this is obviously not the case in our Universe, we must acknowledge that it possesses a certain stability property: small changes in the measurement results require only small adjustments in a theory that predicts these results.

In particular, the same is true as far as space and time measurements are concerned. But it is Lorentz metric that is responsible for them. Consequently, we must ascribe to it the corresponding stability property. We should postulate that small changes in a given Lorentz metric should not produce drastic changes in the (global) space-time structure. In particular, we should postulate that small changes in a given Lorentz metric should not produce closed timelike curves (provided that such curves were absent in the original space-time). The absence of closed timelike curves is evidently related to causality: following such a curve an observer could kill his father before his birth. Therefore, space-times containing no closed timelike curves are justly called *causal* space-times, and any such space-time in which a small perturbation of its Lorentz metric does not produce closed timelike curves is called *stably causal*. If this condition is satisfied, the space-time is not only causal, but also causal with a certain margin of safety, it is not on the verge of violating causality.

And now the surprise. There is a theorem due to Hawking which asserts that space-time is stably causal if and only if it admits a global time.[6] Therefore, if the temporal properties of the world are improved, so as the clocks of local observers indicate the same time flow, then the stability of space and time measurements is automatically guaranteed. To give Hawking's theorem its precise meaning we must determine the meaning of the 'global time' in this context.

[6] S.W. Hawking, 'The Existence of Cosmic Time Functions', *Proc. Roy. Soc.* London A 308, 1968, 433-435.

To extract the geometric meaning from the statement that an observer carries a clock is equivalent to saying that there exists a smooth monotonically increasing function defined along the timelike curve that is the history of this observer in space-time. The real values of this function are interpreted to be indications of the clock carried by the observer. Or, in other words, if an observer carries a clock, its indications associate a real number with each point along observer's history. This defines a real, monotonically increasing function along this history. The global time means that the same function is defined along every timelike curve in space-time. And the Hawking theorem asserts that this is equivalent to the stable causality of space-time.

One more caveat. The existence of global time in the above sense does not presuppose the existence of the universal 'surface of simultaneity': the clocks of all observers indicate the same time flow, but they need not be synchronized. To guarantee such a possibility would require further strengthening causal properties of space-time.[7] But this is another story.

4. *Clocks in the Universe*

So far all our arguments were purely geometric. For instance, we have identified a clock carried by a local observer with a smooth monotonically increasing function along a suitable timelike curve. But any real clock is a physical contraption requiring certain conditions for its construction and functioning. A physical clock is a subsystem of the world, the changes of which could be used to compare them with changes of the environment and to 'monitor' them. If the world were too simple to admit such a subsystem, or too chaotic to allow for its predictable behavior, no clock could be constructed in it. The world in the state of thermodynamic equilibrium, in which only random motions of atoms are possible, is an example of such a 'clockless' system. As noticed by Lee Smolin: 'A world with a clock is then one that is organized to some extent; it is a world somewhere on the boundary between chaos and stasis. The world must be sufficiently dynamical that there is no danger of reaching equilibrium, after which it is chaotic at the microscopic level and static on all larger scales. But it must be organ-

[7] The corresponding space-time should be glabally hyperbolic; see: S.W. Hawking and G.R.S. Ellis, *The Large Scale Structure of Space-Time,* Cambridge University Press, Cambridge, 1973, pp. 206-212.

ized enough that distinct subsystems may be identified that preserve enough order to evolve predictably and simply'.[8]

Every clock, by its very nature, is a cosmological device. 'The point is – writes David Park – that a simple kitchen clock, just as much as the great cosmological models of antiquity, registers the pulse of the universe and keeps time with it, for the same physical laws govern both of them'.[9]

5. *Why Everything Does not Happen at Once*

It was Whitrow who wrote: 'any theory which endeavours to account for time completely ought to explain why it is that everything does not happen at once'.[10] This statement encapsulates intuitions underlying all geometric theorems and their interpretations presented in this paper. In a universe, in which everything happened at once, physics would be trivial as an empty set, and predictions would be impossible by definition. Doing physics presupposes a temporal extension of the world for a dynamics to develop and predictions to be made and verified.

The Universe is a structure which is accessible to us through its various aspects. These aspects, however, are not independent of each other, and some of them are more fundamental than the others. As we have seen, there are strong reasons to think that temporal aspects of the world belong to the most fundamental ones. Some degree of temporality is necessary for the world to be a physical world.

[8] L. Smolin, *The Life of the Cosmos,* Oxford University Press, New York – Oxford, 1997, pp. 287-288.

[9] D. Park, *The Image of Eternity. Roots of Time in the Physical World,* The University of Massachusetts Press, Amherst, 1980, p. 39.

[10] G.J. Whitrow, *The Nature of Time*, Penguin Books, 1975, p. 132.

PREDICTABILITY, DETERMINISM, AND EMERGENCE

JÜRGEN MITTELSTRASS

0. Humans are creatures for whom the future is part of present existence, bounded by uncertainty in many respects, but indispensable for comprehending the present. Immanuel Kant views the 'anticipation of the future' as the 'most decisive proof of man's advantage, in that he is able to prepare for remote objectives in keeping with his destiny'.[1] And for Martin Heidegger, the structure of human existence is future oriented in itself.[2] In one sense this holds for ordinary experience, as reflected in anthropological studies, and in yet another sense it holds for science and leads – in connection with the original Greek idea of order in the physical world – to epistemological analysis. In both areas, predictability is the attempt to deal with the future, and in science – for example in the thesis of the structural identity of explanation and prediction – it is also a crucial criterion of a theory. Predictions serve as both an application of a theory and as its confirmation. The following discussion is limited to addressing the problems connecting to these scientific issues.

1. Problems with predictability in science have been discussed for a very long time. This is particularly so for complex relationships. A classic example is the hole in the ozone layer, or, the effects of chlorofluorocarbons (CFCs) on the high atmosphere ozone layer. In this case, the causal relationships of the chemical reactions are so complex that it is almost impossible to predict their effects. After all, it was difficult enough to explain the mere

[1] I. Kant, 'Mutmasslicher Anfang der Menschengeschichte' (1786), in: *Kant's gesammelte Schriften* VIII, Berlin and Leipzig 1923, p. 113 (Engl. 'Conjectures on the Beginning of Human History', in: H. Reiss [ed.], *Kant: Political Writings*, 2nd ed., Cambridge etc. 1991, p. 225).

[2] M. Heidegger, *Sein und Zeit* (1927), 14th ed., Tuebingen 1977, §§ 67ff. ('Zeitlichkeit und Alltaeglichkeit').

occurrence of the effect. Just as well, it is a common fact that small causes can have large, unpredictable effects. Ice ages, for example, according to recent scientific research, are caused by a relatively minor cooling down in the earth's atmosphere. This in turn, is caused by a decreased intensity in the rays of the sun, which results from peculiarities of the earth's revolving around the sun, in particular its varying eccentricity as well as variations in its orientation and the gradient of the earth's axis. The crucial point is that this trifling cooling down leads to a change in flow in the North Atlantic. In particular, the warm flow, which comes to the surface near Iceland and is responsible for the warm climate in Europe, is diverted. This leads to a much harsher climate in the north, which in turn contributes toward cooling at the global dimension.[3] Thus small changes in the conditions cause, in this case, considerable changes in the state of the system as a whole.

Another example is related to Max Planck's (epistemologically problematic) exploration of free will, which has recently become relevant again for brain science. Embarking from the concept of causal universality, i.e. the assumption of causal closure of the world, Planck argues that the will is also causally determined, although mental events, e.g. thoughts, are unpredictable – even for an ideal observer – due to their manifold dependencies. For Planck, this is also relevant for the relations between a willing and a perceiving self (the ideal observer): 'Each new observation (...) gives rise to a new motive, and the recognition of this motive in turn creates a new situation. The series is infinite, and since the observed person (the willing ego) owes no obedience to the observer (the percipient ego), we shall never be able to claim with certainty that the eventual decision must be in the sense of the observer's latest discovery'.[4] This has, following Planck, no bearing on the continued validity of a causal law.

2. On this topic, the most commonly discussed example is chance in quantum mechanics. Quantum mechanics imposes serious limitations on the predictability of events. The central principle of the theory is 'Schroedinger's equation', which serves to determine the 'state function' or 'wave function' of a quantum system. The state function is generally taken to provide a complete description of quantum systems; no properties can

[3] See W.S. Broecker and G.H. Denton, 'Ursachen der Vereisungszyklen', *Spektrum der Wissenschaft* 3, 1990, pp. 88-98.

[4] M. Planck, *Vom Wesen der Willensfreiheit*, 2nd ed., Leipzig 1937, p. 18 (Engl. *The Universe in the Light of Modern Physics*, 2nd ed. [with a section on Free Will], London 1937, p. 101).

be attributed to such a system beyond the ones expressed in terms of the state function. Schroedinger's equation determines the time development of the state function unambiguously. In this sense, quantum mechanics is a *deterministic* theory.

However, apparently irreducible chance elements enter when it comes to predicting the values of observable quantities. The measurement process in quantum mechanics is described as the coupling of the quantum system to a particular measuring apparatus. Schroedinger's equation yields, then, a range of possible measuring values of the quantity in question, each of these values being labelled with a probability estimate. That is, Schroedinger's equation only provides a probability distribution and does not anticipate particular observable events. Quantum mechanics is extended to actual measuring values by adding the so-called 'projection postulate'. This postulate is independent of Schroedinger's equation and says that one of the possible measuring values is assumed in actuality. The spectrum of possible values collapses into the one value that is obtained in the measurement. In repeated measurements of the same kind, the relative frequencies of the values coincide with the probability estimates supplied by Schroedinger's equation.

The salient point is that, according to present lights, this collapse of the state function, i.e., the selection of the actual measuring value from the range of possibilities is a genuinely *indeterministic* process whose outcome cannot be predicted on any basis whatsoever. These obstacles to prediction, as they become manifest in quantum mechanics, have nothing to do with the ignorance of the prevailing initial conditions. Given a complete description of the quantum state, chance fluctuations at the level of observables will yet occur. Quantum mechanics involves in-principle limitations of predictability to the effect that, for instance, it is objectively indeterminate when a given radioactive nucleus will decay. Such limitations are not merely epistemic constraints, but rather represent an ontological indeterminateness.

Heisenberg's so-called indeterminacy relations are a consequence of Schroedinger's equation, although historically they were formulated independent of this equation and prior to its enunciation. The Heisenberg relations place severe limitations on the simultaneous measurement of what is called 'incompatible' or 'incommensurable' quantities like position or momentum or spin values in different directions. The more precise one of the quantities is evaluated, the more room is left for the other one. Like the constraints mentioned before, the limitations set by the Heisenberg relations have nothing to do with practical impediments to increasing measurement accuracy that might overcome by improved techniques. Rather, the relations express limitations set by the laws of nature themselves.

Heisenberg's indeterminacy relations entail serious restrictions of the prediction of future quantum states. For ease of illustration consider the following spin measurements. Spin states are quantized; they possess only two possible values in each direction, namely, 'spin up' or 'spin down'. A beam of electrons can be 'spin-polarized' by sending the particles through a suitably shaped magnetic field (a Stern-Gerlach apparatus). That is, the spin of all electrons in, say, x-direction after exiting from the setup is, say, 'up'. This result can be confirmed by a second measurement of the same quantity performed directly after the first. 100% of the electrons come out 'spin up' in the x-direction. Let the beam then pass through the same setup but now measuring the spin values in the y-direction, perpendicular to x. The outcome is that one half of the beam exhibits 'spin up' and the other half 'spin down'. If the beam is finally sent through the apparatus this time oriented again in x-direction, the perplexing result is that 50% of the electrons are registered 'spin up' and 'spin down', respectively.

Correspondingly, the first measurement, in spite of its quite unambiguous result, cannot be utilized for a prediction once a measurement of an incompatible quantity has been carried out. Again, this is a matter of principle. There is no way of anticipating the joint values of incompatible quantities below the threshold set by the Heisenberg relations. As a result, inherent limitations prevent us from predicting the future states of such quantities.

This element of genuine, irreducible chance troubled Albert Einstein very much. Einstein accepted statistical accounts if they could be viewed as growing out of incomplete knowledge of the relevant conditions and states. Quantum mechanics differed from all other statistical theories in physics in that the invocation of probability could not be attributed to human ignorance. Einstein's commitment to a determinist world was his chief reason for dissenting from quantum mechanics. As he wrote to Max Born, he found the idea 'unbearable' that an electron decides on its own in which direction to move. If this turned out to be true he preferred to be an employee in a gambling casino rather than a physicist.[5] In the same vein, Einstein told Born that quantum mechanics does not bring us closer to God's mystery. After all, God does not throw dice.[6] This episode bears witness to the fact

[5] Einstein to Born (April 29, 1924), in: *Albert Einstein – Hedwig und Max Born: Briefwechsel 1916-1955*, Munich 1969, p. 118.

that in-principle constraints on predictability represent a serious deviation from the notion of *Laplace's demon* which is the core element of the traditional, ignorance-focused account of chance and probability.

To repeat once more: Current wisdom holds there are fundamental processes in the quantum world that inhibit randomness, which implies general limits of predictability. Nevertheless this is by and large irrelevant to macroscopic phenomena; with large numbers of atoms the uncertainties average themselves out. This, in turn, brings us to the fundamental question of the relationship between *determinism* and *predictability*.

3. Talking about the limits of predictability in principle immediately poses questions for a *deterministic world*. This has been clear to Max Planck, leading to the insight that the classical dictum 'an event is causally conditioned, if it can be predicted with certainty'[7] cannot be maintained, moreover, one is forced 'to acknowledge the following sentence as a given fact: In no circumstance is it possible to predict a physical event with exactness'.[8] In a similar vein, several years before (1927), Werner Heisenberg claimed that, 'in principle', quantum mechanics has the effect that, 'the law of causality is in a sense unfounded. Since one can never know precisely the initial conditions, one can never calculate the mechanical course of events. (...) Concerning the sharp version of the law of causality: If we know the present, we can calculate the future – it is not the consequent, but the antecedent that is wrong'.[9] This, however, is not the last word on the possibility of a deterministic world. It is rather necessary to separate the concepts of determinism and predictability from each other; determinism understood here (following J. Earman) as the thesis that, if two possible worlds are identical at a given point in time, then they are identical at every point in time.[10] This does not exclude hindrances to predictability for a given state of affairs and deterministic development. The thesis is: *Even in a*

[6] Einstein to Born (December 4, 1926), *ibid.*, p. 129.

[7] M. Planck, 'Die Kausalität in der Natur' (1932), in: M. Planck, *Vortraege und Erinnerungen*, 5th ed., Stuttgart 1949, p. 252.

[8] *Ibid.*, p. 253.

[9] W. Heisenberg, 'Über die Grundprinzipien der Quantenmechanik', *Forschungen und Fortschritte* 83 (1927), p. 83 (= W. Heisenberg, *Gesammelte Werke / Collected Works*, vol. 1, Munich and Zurich 1984, p. 21).

[10] See J. Earman, *A Primer on Determinism*, Dordrecht etc. 1986, p. 14 ('if two worlds agree for all times on the values of the conditioning magnitudes and if they agree at any instant on the values of the other magnitudes, then they agree at any other instant').

deterministic world there are limits of predictability.

Two reasons can be given in support of this. First, *deterministic chaos*. This refers to the strong dependence of a system's states of affairs on the magnitude of defined parameters. Since the magnitude of these parameters can never be known, the prediction of a system's states of affairs is bound by uncertainty, which translates into a range of different developments in chaotic systems. Unpredictability as a result of chaos is not limited to complex systems, rather, it can also occur in simple systems that only consist of a few elements. For example, two coupled pendulums constitute a simple system, the relevant laws of which have been known for centuries. But it has only recently become clear that, within such an arrangement, in a distinct range of initial conditions – namely system stimulations of medium strength – there can be chaotic and unpredictable oscillations. Another example, already introduced in the beginning, is meteorology, which was the original impulse for studying chaotic effects in dissipative systems. In a well-known metaphor: even the flapping of a butterfly's wings can crucially effect the convection currents in the earth's atmosphere and, hence, meteorological developments ('butterfly effect').[11] The reliability of weather forecasts is not only constrained by *practical* limits but also by limits *in principle*. These occur even though the underlying laws are known and of a deterministic nature.

More generally and again using the example of weather forecasting, this can be formulated as follows:[12] it is possible to know the exact equations of motion for a system, without being able to predict the evolution in time of this system. Although meteorological developments (as it is generally understood) can be completely described by thermodynamic equations, this is of little help. Because all observations are always finitely accurate, the future behaviour cannot be predicted using these equations. Though weather can be predicted in the short run; the chaotic effects described here will still appear in the middle to long run. It is important to see that it is not the system itself that behaves chaotically; its development is, quite the contrary, strictly deterministic. A chaos exists only for us, not for the thing

[11] Cf. H.G. Schuster, *Deterministic Chaos: An Introduction*, Weinheim 1984, p. 2.

[12] This is also the conclusion of a longer argument in: M. Carrier and J. Mittelstrass, *Mind, Brain, Behavior: The Mind-Body Problem and the Philosophy of Psychology*, Berlin and New York 1991, p. 262. Cf. M. Carrier, 'Chaostheorie', in: J. Mittelstrass (ed.), *Enzyklopaedie Philosophie und Wissenschaftstheorie*, vol. 2, 2nd ed., Stuttgart and Weimar 2005, pp. 40-43.

being studied; it results from the imprecision of our knowledge of the initial conditions. But this means that there is an *epistemological limit* that occurs in the phenomenon of deterministic chaos. Although a system is, in fact, strictly deterministic and can be completely understood according to certain laws, we are not in a position to describe the behaviour of this system, despite our precise knowledge of these laws.

Epistemologically speaking the chaos is a *supervenient* characteristic.[13] A characteristic *s* is supervenient to a set of physical characteristics *p*, if (1) *s* is not of a concrete-physical nature, that is, *s* can obtain in physically different systems, and if (2) differences in *s* always coincide with differences in *p* (although not *vice versa*). The occurrence or non-occurrence of chaos always depends on the physical differences in the system.

The second reason is the problem of a *Laplace's demon*. This label (credited to E.H. Du Bois-Reymond)[14] refers to a fictitious superhuman intelligence, which – under the assumption of a stable, closed and all-determined system typical for a mechanistic worldview – knows of all initial conditions of all possible movements and thus can predict the location of any particle for every point in time. Now (as has already been mentioned), quantum mechanical systems – in contrast to relativistic physics, where differential equations describe deterministic systems with regards to their state variables – are non-deterministic with regard to conjugate variables such as position and momentum. Rather, they are statistic, i.e. incalculable even by Laplace's demon – an implication confirmed by recent developments in physics.

There is yet another reason why Laplace's demon is unable to handle the problem of predictability, even under the assumption of deterministic structures.[15] Such a demon would himself be part of the world which he seeks to predict. This situation inhibits *self-reference*: the observing system or measurement device is itself part of the system whose development is being predicted. In other words, predictability in a Laplace's demon situation demands measurability of a system state 'from within'. This, in turn,

[13] Cf. P. Hoyningen-Huene, 'supervenient / Supervenienz', in: J. Mittelstrass (ed.), *Enzyklopaedie Philosophie und Wissenschaftstheorie*, vol. 4, Stuttgart and Weimar 1996, pp. 144-145.

[14] 'Über die Grenzen des Naturerkennens' (1872), in: E. Du Bois-Reymond, *Vortraege über Philosophie und Gesellschaft* (ed. S. Wollgast), Hamburg 1974, pp. 56-57.

[15] For the following see Th. Breuer, *Quantenmechanik: Ein Fall fuer Goedel?*, Heidelberg etc. 1997, pp. 7-21; 'Limits to Self-Observation', in: M. Carrier *et al.* (eds.), *Science at Century's End: Philosophical Questions on the Progress and Limits of Science*, Pittsburgh Pa. 2000, pp. 135-149.

demands, (1) that any object state is connected to the state of an apparatus, hence there can be no object states remaining (though it is possible that apparatus states can exist without a corresponding object state), and (2) that there are no two object states that correspond with the same apparatus state (while conversely there can be two apparatus states which correspond to the same object state).[16] To measure each object state with exactness, it must correspond to at least one apparatus state. This implies first and foremost that there are at least as many apparatus states as there are object states. However, the assumption of the *inner observer* (a demonic situation) implies that there are more object states than apparatus states. An inner observer is indeed part of an object, such that for the inner observer the apparatus states are a proper subset of the object states.

These conditions contradict each other. The demand for exact measurement implies that the apparatus has at least as many states as the object. The condition of the inner observer says that the object has more states than the apparatus. These two conditions cannot hold at the same time. And this is a strong argument for the separation of predictability and determinism. Both arguments, the chaos argument and the argument of the inner observer, make it clear that there can be deep or even basic problems of prediction even in a deterministic framework; hence determinism and unpredictability are not mutually exclusive.

4. Laplace's demon has lost its demonic character in this context; he has become an observing scientist. Thus it has been rightly said: 'In fact, most of the contributors to the debate, having paid lip service to Laplace, almost unnoticeably substitute for his demon a human observer. They thereby reduce determinism to predictability, i.e., the question whether an actual observer, a biologist or a physicist, is able to predict future events. This reduction of Laplacian determinism to actual predictability is a drastic step. On the one hand, it brings the question from philosophical clouds down to earth, where one may hope to find an answer. On the other, it is reduced to a technical question about the state of affairs in the relevant science'.[17] This is also the case with K.R. Popper.

For Popper, who by and large identified determinism and predictability with each other (determinism = predictability with a defined level of exactness, which depends on the degree of knowledge about the initial condi-

[16] In Th. Breuer's presentation this is expressed by the subjectivity of the picture ('Can a picture contain a full and precise picture of itself?'), 'Limits to Self-Observation', p. 135.

[17] N.G. van Kampen, 'Determinism and Predictability', *Synthese* 89 (1991), p. 275.

tions),[18] *scientific determinism* is 'the doctrine that the state of any closed physical system at any given future instant of time can be predicted, even from within the system, with any specified degree of precision, by deducing the prediction from theories, in conjunction with initial conditions whose required degree of precision can always be calculated (in accordance with the principle of accountability) if the prediction task is given'.[19] In this context, Laplace's demon is construed as a disembodied spirit; he is transformed into a 'super-scientist': 'The demon, like a human scientist, must *not* be assumed to be able *to ascertain initial conditions with absolute mathematical precision*; like a human scientist, he will have to be content with a finite degree of precision'.[20] Naturally, this leaves room also for deterministic conceptions.

Popper's critique of determinism in natural science and philosophy employs arguments not only from quantum mechanics, but also from classical physics. He argues that Newtonian mechanics, which is deterministic by conception, is unable to determine initial conditions with the precision necessary for prediction ('principle of accountability'). More generally, according to Popper, the growth of theoretical knowledge is not predictable in principle, which also hints at an *indeterminism* – which could be used, for example, as a solution for mind-body problems.

5. Let me refer in a final part to the concept of *emergence*. Emergence says that it is impossible to use characteristics of elements and the interrelations between these to describe characteristics of ensembles or make predictions about them.[21] Thus a common formula says this: the whole is more than its parts.[22] According to the *emergence thesis*, the world is a levelled structure of hierarchical organised systems, where the characteristics of higher-level systems are by and large fixed by the characteristics of their respective subsystems, yet at the same time essentially different. Different characteristics and processes occur in the respective levels. As well, a weak and a strong emergence thesis can be distinguished from one another.

The core element of the strong emergence thesis is a non-derivability-

[18] Cf. K.R. Popper, 'Indeterminism in Quantum Physics and in Classical Physics', *The British Journal for the Philosophy of Science* 1 (1951), pp. 117-133.

[19] K.R. Popper, *The Open Universe: An Argument for Indeterminism*, Totowa 1982, p. 36.

[20] *Ibid.*, p. 34. Cf. J. Earman, *A Primer on Determinism*, pp. 8-10.

[21] For the followings see M. Carrier, 'emergent / Emergenz', in: J. Mittelstrass (ed.), *Enzyklopaedie Philosophie und Wissenschaftstheorie*, vol. 2, 2nd ed., Stuttgart and Weimar 2005, pp. 313-314.

[22] Cf. K. Lorenz, 'Teil und Ganzes', in: J. Mittelstrass (ed.), *Enzyklopaedie Philosophie und Wissenschaftstheorie*, vol. 4, Stuttgart and Weimar 1996, pp. 225-228.

or non-explainability hypothesis of the system characteristics shaped from the characteristics of the system components. An emergent characteristic is non-derivable; its occurrence is in this sense unexpected and unpredictable. Weak emergence is limited to the difference of the characteristics of systems and system components and is compatible with the theoretical explainability of the system characteristics. Weak emergence is essentially a phenomenon of complexity.

The classic rendering of strong emergence is credited to Ch.D. Broad.[23] Broad's motivation was to provide a suitable interpretation of living organism. He intended to depict organisms neither as mere machines nor as being fuelled by an exceptional vital force. This neo-vitalist view was first and foremost endorsed by H. Driesch,[24] who maintained that beings are fitted with 'entelechy', i.e. with purposeful biological powers. Broad was searching for a third way between the mechanistic and the vitalist view on life. The emergence thesis was intended to create this path. Emergent characteristics of ensembles were intended to be roughly defined by the divergent characteristics of their components, yet it was not intended to explain the former on that basis.

Strong emergence is characterised through the following conditions: (1) The condition of *qualitative difference*. This condition applies the emergence thesis to those characteristics of ensembles which differ profoundly from the characteristics of their components. (2) The condition of *characteristic determination*. This condition says that the characteristics of the components are sufficient to let the specific characteristic emerge; emergence is not depended upon further factors. (3) The condition of the *principal gap of explanation*. This condition implies that it is actually impossible to explain the characteristics of ensembles through the characteristics of their components, including their interrelations. – Incidentally, the existence of strong emergent characteristics in this sense is heavily disputed. The only candidates in the running at the moment are currently *phenomenal* characteristics.[25] The point here would be, that in a given neurophysiological arrangement, the occur-

[23] Ch. D. Broad, *The Mind and its Place in Nature*, London 1925.

[24] H. Driesch, *The Science and Philosophy of Organism*, vols. 1-2, London 1908, 2nd ed., London 1929.

[25] See A. Stephan, 'Phaenomenale Eigenschaften, phaenomenale Begriffe und die Grenzen Reduktiver Erklaerung', in: W. Hogrebe and J. Bromand (eds.), *Grenzen und Grenzueberschreitungen* (XIX. Deutscher Kongress fuer Philosophie, Bonn, 23.-27. September 2002). *Vortraege und Kolloquien*, Berlin 2004, pp. 404-416.

rence of qualitative experiences (e.g. blue, the sound of trumpets etc.) in a system would be non-derivable and unpredictable.

Concerning predictability, it is particularly the *temporal* aspect of the emergence thesis which is of interest, i.e. for ensemble characteristics that occur in developments. Limits of reductability (of the whole to its parts) figure here as limits of explanation and predictability. This temporal novelty is described by the concept of *creative advance of nature*. It is endorsed by Popper and Eccles, among others.[26]

6. To sum up: determinism does not imply predictability, and unpredictability does not imply non-determinism. In fact, there is unpredictability in a deterministic world, and unpredictability permits deterministic worlds. This has been illustrated with the discussion of the concepts of deterministic chaos, Laplace's demon, who becomes stripped of all his demonic characteristics, and emergence. Besides, one could not simplify matters by distinguishing (as has been proposed) between ontological determinism and epistemic unpredictability. First, such a distinction is epistemic in itself and second, it merely expresses that the concepts of determinism and predictability do not belong to the same (semantic) level, or even mean the same thing. Predictability, not determinism, is the problem (in some areas). Dealing with unpredictability in the right way is the challenge – in science as well as in ordinary life.

I would like to thank Martin Carrier (University of Bielefeld) for his assistance, particularly in the section on quantum physics.

[26] K.R. Popper and J.C. Eccles, *The Self and Its Brain*, New York etc. 1977, pp. 22-35. Cf. also M. Čapek, *The Philosophical Impact of Contemporary Physics*, Princeton N.J. etc. 1961, pp. 333ff.

PREDICTABILITY: PROPHECY, PROGNOSIS AND PREDICTION. A STUDY IN NEUROEDUCATION

ANTONIO M. BATTRO

1. The Cognitive Space of Predictability

I would like to interpret predictability as a dynamical process in a cognitive space defined by three independent parameters: prophecy, prognosis and prediction. Three cognitive functions are engaged in this process: projection, anticipation and modelling which respectively apply to possible, actual and constructed worlds.

Parameters	**Functions**	**Worlds**
prophecy	projection	possible
prognosis	anticipation	actual
prediction	modelling	constructed

I will now try to describe this 3x3 table in order to display the framework of our following discussion on predictability and of the new concept of metaprediction.

Prophecy

It may seem odd to refer to prophets and prophecies in this scientific context but from the historical view of the human mind it is impossible not to consider the weight of this venerable capacity of announcing the future, denouncing the wrong paths and proposing favorable changes. Our culture is rich in prophecies old and new of every kind which are sometimes the motor of remarkable transformations in our societies. Science also has its prophetic dimension and we need to take it into account if we want to understand what predictability in science means. At particularly hard times 'scientific prophets' are most needed as John Paul II in his address of 12 Nov. 1983 to the Pontifical Academy gathered to discuss *The knowledge that builds peace* emphasized:

> Unarmed prophets have been the object of derision in every age, especially on the part of shrewed politicians, the supporters of power. But today must not our civilisation recognise that humanity has need of them?...the scientists of the whole world ought to be united in a common readiness to disarm science and to form a providential force for peace (*Papal Addresses,* 2003, p. 260).

Scientific prophecy is based on *projections*. This means that scientists may project into a possible future their personal views and wishes in order to produce a desiderable change. For instance, some fifty years ago some pioneers of artificial intelligence and computer science imagined a digital scenario for education, but this projection was received with wide skepticism and even strong opposition. Computers at that time were expensive and unaccessible to teachers and students, but good projections might work as 'powerful ideas' and may induce a radical change against all prejudices (Minsky, 1986, Papert, 1980, 1993). A decade ago Nicholas Negroponte in his book *Being Digital* (1995) described the coming of age of the digital era in education announced by the 'prophets' and today he leads an ambitious – and prophetic – program to empower millions of children with computers and communications, the OLPC, the one laptop per child initiative, a new projection towards a profound change in a globalized education (http://laptop.org/).

Something similar, and closely related to the computer revolution, is now unfolding in the new field of neuroeducation. We can make also the (prophetic) projection that in a generation from now the neurocognitive sciences will provide a radical new basis for learning and teaching and will open not only cognitive but also ethical issues of great impact (Sheridan, Zinchenko & Gardner, 2005, Battro, Fischer & Léna, in press). I will describe some landmarks in this prophetic educational endevour that has engaged many of us around the world and is leading towards a new view of the learning and teaching brains.

Projections are about *possible (future) worlds*. This means that they go beyond what is given but they are not fiction, even less science fiction. As Nelson Goodman said 'all possible worlds lie within the actual one' (Goodman, 1979). Projections are forecasts about something that will possibly happen given some necessary conditions, they are not dreams or fantasies. Projections, of course, can be seductive and tempting but can also lead astray. And most important, the ethical component is an essential part of any scientific projection and should be made explicit.

Prognosis

The term was coined by the physicians of antiquity and is still used in medicine, where prognosis and therapy are closely related in everyday medical practice. Physicians make a prognosis of the disease diagnosed in the patient. This interplay of the universality of the disease and the individuality of the suffering organism is the reason why medicine is science and art at the same time. Many disciplines share this double condition, for example in education the general sciences of teaching and learning are embodied in individual cultures and values. Neuroeducation is also under this intrinsic tension between the universal and the particular, a tension that is an important source of progress indeed.

Prognosis is based on *anticipations*. In the same way as the physician anticipates the unfolding in time of the disease and takes a number of decisions for the best treatment of the patient the scientist also anticipates the course of events and prepares the conditions or working scenarios that will help to unfold the discoveries and to control, if possible, their undesiderable consequences.

Anticipations are made in the *real (actual) world*. They should not be confused with projections which deal with possible (future) worlds. Anticipations are effective innovations, they are at the cutting edge of research, at the frontier of knowledge but never leave the real world, they never go beyond the actually given but they lead the transformation. Anticipations can fit and be successful, but they also can be premature or fail altogether.

Prediction

The capacity to predict is common to many animal species, it is a condition of survival (Llinás, 2002, Dehaene *et al.*, 2005). We need to better understand the neurocognitive processes involved in our capacity to predict. Science is the result of a consistent and permanent construction of predictions that can be traced in the evolution of the brain and in the development of the human mind. Science can be taught to children in schools because they are already able to make predictions, eager to test the validity of their own hypotheses and pleased to teach other children (Piaget, 1949, Charpak, Léna & Queré, 2005). This incredible teaching capacity is unique to the human species and can be analyzed with the conceptual tools and the technology of the modern neurocognitive sciences (Strauss, 2005, Battro, 2006).

Prediction is based on *representations or models*. We must recognize that the use of brain models and the interpretation of brain images have opened a whole new field of research. We now have the resources to explore the neuronal intimacy of several cognitive processes, and in some cases we can even predict the expected behavior from the brain images which represent the neuronal activy involved in the mental process. I call 'metaprediction' this prediction on predictions.

Models are *miniature or constructed worlds* that we can handle and simulate in our minds or in a computer. Models must be tested and when they fail the model becomes invalid. Karl Popper has also proposed that even when the model is verified it is still impossible to assert its universal and necessary validity because it could be falsified in the future (Popper, 1959). Models are not 'the' world but a – fragile and reduced – 'representation' of the world. The great discovery is that this formal constraint empowers the model to predict an event. We can, for example, identify the different neurocognitive processes that are involved in making Aristotelian or Newtonian predictions and this recent discovery implies that we can detect different models of the physical world in our brains (Fugelsang & Dunbar, 2005).

2. A Survey on Predictability

I propose now to discuss the three cognitive dimensions of predictability from the point of view of the neurocognitive and educational sciences.

a) Prophecy, Projections and Possible Worlds

Seymour Papert gives a lively example of a prophetic scientific vision. A mathematician by training he became a close collaborator of the psychologist and epistemologist Jean Piaget in Geneva and of the computer scientist Marvin Minsky at MIT, where they co-directed the famous Laboratory of Artificial Intelligence in the sixties. This double collaboration was substantial to introduce the computer in the schools around the world. Papert has described the elaboration of his prophetic view on the future of digital education with the following words (Papert, 1993, pp. 34-34):

> It was pure play. We were finding out what could be done with a computer, and anything interesting was worthwhile. Nobody yet

> knew enough to decree that some things were more serious than others. We were like infants discovering the world.
>
> It was in this situation that I thought about computers and children. I was playing like a child and experiencing a volcanic explosion of creativity. Why couldn't the computer give a child the same kind of experience? Why couldn't a child play like me? What would have to be done to make this possible?
>
> These questions launched me on a new quest guided by the Robin Hood-like idea of stealing technology form the lords of the laboratories and given it to the children of the world. A first step in the quest was to recognize that one of the sources of the technologist's power was the veil of esoteric mystery woven around the idea of programming...I saw the need to make computer languages that could be 'vulgarized' – made available to ordinary people and especially children.

This quotation expresses the profound personal engagement that fuels any worthy projection in the sciences and the origin of the prophetic vision that, in the case of the information and communication technologies, has changed the world of education. Papert invented the computer language called Logo that became a powerful tool in the hands of thousands of teachers and students. For those who were active in the first wave of the digital education it is clear that we are now repeating a similar cycle in the field of neuroeducation. We also feel the need to 'take away' the brain imaging technologies from the laboratories and bring them into the schools. In order to do that, first, we must have portable, reliable, simple and low cost brain imaging equipment (as was the case with the first PCs in the seventies) and second, we should train the 'neuroeducators' in the creative use of these machines. We are still quite far away from this prophetic scenario but we must prepare the field of neuroeducation to be able to cope with it, when the time will come. Only one generation ago very few were convinced by Papert's digital prophecies, similarly we shouldn't be disturbed by the skeptics of today about the future of neuroeducation.

b) *Prognosis, Anticipations and Actual Worlds*

Muchas veces me dicen que me anticipo a propiciar cosas que sólo serán posibles de aquí a 30 a 40 años. Pero eso no es exacto, porque preconizo lo que es actual y urgente, que ya existe en los países más adelantados, mientras mis contradictores no lo saben porque están 30 a 50 años atrasados y lo ignoran.

(Bernardo A. Houssay, 1967)

Anticipation is the result of a correct diagnosis of the current state of scientific research and a fair prognosis of its intrinsic development. It should not be confused with a projection, which is a forecast of future improvements. In other words, while projection invents the future, anticipation unfolds the present, they are two different – and complementary – cognitive functions. The following text is a strong statement about the value of anticipation and belongs to a master Argentine scientist, Bernardo A. Houssay, Nobel Prize in medicine (1948), the first to become a member of the Pontifical Academy of Sciences from Latin America (1936), a leading figure in promoting science in developing countries. He understood anticipation as an urgent scientific mission in his own time and country (Delofeu & Foglia, 1982):

> Several times I have been told that I anticipate and promote things that will only be possible in 30 or 40 years from now. But this is not correct because I support what is actual and urgent, what exists already in the most advanced countries, while my contradictors don't know that because they are 30 or 50 years behind and they ignore it.

In the same spirit Kim Sheridan, Elena Zinchenko and Howard Gardner (2005) anticipate some crucial ethical issues in neuroeducation when brain imaging techniques will become standard in the school practice. They take the hypothetical example of a teacher, Ms. E and a student Daniel with some minor learning disablity:

> In the first scenario we see Ms. E confronted with a variety of traditional assessments (e.g., standardized tests scores, last years' grades, current work) and a type of neurological report that will likely become standard in the near future. One of her core values as an educator is to help each student develop to the best of his or her potential, including seeking remediation for any learning disabilities. In an ideal case, a neurological evaluation yields a clear diagnosis to which an empirically valid remediation is yoked. However, it is likely that there will be many more like this hypothetical one: a report of atypical processing for which there is neither a clear diagnosis nor remediation.
>
> In assessing the fMRI report, Ms. E is expected to don a hat for which she is inadequately trained. In the face of her lack of expertise and the fMRI report's ambiguity, the report seems to reveal something 'true' about Daniel's functioning. She allows the biological finding to trump her observations as a teacher and Daniel's hitherto adequate performance in class. Drawing upon her classroom

> observations and educational training, Ms. E may have given Daniel a positive report for his current progress and perhaps worked out some in-class or at-home strategies for his minor attention and reading issues. However, faced with this picture of his brain function, she feels out of her depth and considers remediation strategies that may not be appropriate for Daniel's needs'.

The use of brain imaging technology in the practice of remediation and special intervention in this hypothetical case anticipates what might become common practice tomorrow by the increasing interaction between neuroscientists and neuroeducators in many other fields. We already have the successful example of the great change produced by the use of advanced technology in special education and rehabilitation (computer prostheses, cochlear implants, special software, robotics, etc). Something similar can be anticipated in neuroeducation when the intimacies of fundamental neurocognitive processes will become the targets of teaching and learning practices in all the disciplines, well beyond the current demands of remediation and cognitive enhancement in special cases. We expect that a new generation of 'neuroeducators' will be trained in the most diverse fields of the arts and sciences. We can expect great changes in the way we teach and learn because of this interaction.

c) Prediction, Modelling and the Constructed Worlds

> *The capacity to predict is most likely the ultimate brain function*
> (Rodolfo Llinás, 2002)

A historian of science would identify different steps of the path of predictability showing in some cases a strong weight of the prophetic dimension at the beginning of a specific scientific quest and an increasing weight of the prognosis dimension later in time. Then, at some point of the evolution of the sciences and technologies involved in our quest the new dimension of prediction will start to unfold and the path of predictability will show a radical change because of the possibility to make specific predictions in the new fields by modelling and testing. And again the new discoveries will start a new cycle of prophecies, prognoses and predictions, a quest without end.

A remarkable and most important extension of the concept of predictability arises from the neurosciences, in particular with the new possibility to make 'metapredictions' i.e. predictions about predictions during a cognitive task. A robust demostration about the possibility to infer behavior from functional brain images was provided by Stanislas Dehaene and colleagues (Dehaene *et al.*, 1998). In this experiment the subjects were asked to press a key with the right or the left thumb to decide whether digits presented visually were larger or smaller than 5. They used one-line brain activation measurements to predict the subject's decision on number comparison, reversing the standard practice which goes from the known behavior to the specific brain activity. In fact, the observers are doing a kind a 'reverse neurology', they were making 'brain predictions' on the motor behavior (*prediction 1*: left and right cortical activations predict respectively the right and left finger movements) to be applied upon the 'mental predictions' on the arithmetics (*prediction 2*: larger or smaller than 5). The whole experiment is about metapredictions: the observer predicts that a particular brain activation predicts the mental decision of the subject. Pressing the key is only the final step of a complex chain of brain events like identifying the name or Arabic symbol, interpreting it as a quantity, making a comparison with the given number (5) and making a quick decision (larger or smaller). This chain of brain events follows a well known space-time path in the cortex (Dehaene, 2006) measured in several fixed steps of milliseconds. Moreover, 'reverse neurology' opens the possiblity to predict not only some overt and well controlled behaviors – as in this experiment on number comparison – but also some covert behaviors and intimate mental processes (internal speech, emotions, visual imagery, etc) from their neuronal activity pattern. We must agree with the authors that this possibility raises important practical and ethical questions. In neuroeducation, we can imagine, for instance, that we will someday explore the learning process of the students by 'looking into their brains' in addition to the evaluation of their performance on a standard test.

A first step in this direction has already been taken, and is the following. It is well known that students have trouble overcoming naïve explanations about the movement of bodies, for instance. Andy di Sessa (1982) some decades ago showed with the aid of computers the amazing difficulty of students to interact in a Newtonian world where forces correlate to velocity and not to position as in Aristotelian mechanics. Most students have a preferred set of concepts (called phenomenological primitives by di Sessa) that are in contradiction with what they have learned in the physics class but they still use the 'impetus' idea that objects move in the direction you push them. In order to use the Newtonian model to correctly predict the movement of an object they must 'unlearn' the common intuitions of the Aristotelian physics. In a strong sense successful education implies many 'conceptual changes' of this type. The problem is that naïve students tenaciously hold on to their preferred model. In the same vein Jonathan Fugelsang and Kevin Dunbar (2005) have recently studied the brain images of students during a fMRI experiment on their preferred theories in physics. They have tested two groups of subjects, physics students and non-physics students looking at 'Newtonian' and 'naïve' movies where balls of different sizes fall at equal or different rates in a frictionless environment. If the balls fall as they expected they must press one key, if not the other key. For the naïve students the Newtonian film is erroneous, for the physics students (who made the correct conceptual change) it is the naïve film which is erroneous. In other words, the non-physics students predict that the larger ball will fall faster than the smaller one (as in the naïve film) while the physics students predict that both balls fall at the same rate (as in the Newtonian film). The difficulty in teaching physics is related to the fact that sometimes the old and naïve model persists in the student's mind even when the intended conceptual change appears to have taken place. The fMRI records an increased activation in the Supplementary Motor Area and in the Anterior Cingulate that may 'inhibit' those data that become inconsistent with the student's preferred theory in both groups. It is known that the Anterior Cingulate cortex is related to error detection and in general the medial-frontal cortex is activated by existing conceptual representations. This experiment shows that the physics students inhibit the counter-Newtonian data and that the non-physics students inhibit the counter-intuitive data. The metaprediction in this experiment is that the conceptual change should be reflected in the brain where the old model must be inhibited because it is detected as an error in order to give place to the new model. This is only a first – and indirect – method to test the value of metapredictions in neuroeducation but it opens a whole new horizon in the quest of predictability in science.

REFERENCES

Battro, A.M., Fischer, K.W. & Léna, P.J. (in press), *The educated brain*, Cambridge: Cambridge University Press.

Charpak, G., Léna, P. & Queré, Y. (2005), *L'enfant et la science. L'aventure de la main à la pâte,* Paris: Odile Jacob.

Dehaene, S., Le Clec'H, G., Cohen, L., Poline, J.-B., van de Moortele, P.-F. & le Bihan, D. (1998), 'Inferring behavior from functional brain images', *Nature Neuroscience*, 1, 7, 549-550.

Dehaene, S., Duhamel, J.R., Hauser, M. & Rizzolatti, G. (eds.)(2005), *From monkey brain to human brain*, Cambridge, MA: MIT Press.

Dehaene, S. (2006), *Psychologie cognitive expérimentale. Vers une science de la vie mentale,* Leçon inaugurale au Collège de France, 27 avril 2006.

Di Sessa, A. (1982), 'Unlearning aristotelian physics; A study of knowledge-based learning', *Cognitive Science*, 6, 37-75.

Foglia, V. & Delofeu, V. (eds.)(1981), *Bernardo A. Houssay. Su vida y su obra 1887-1971,* Buenos Aires: Academia Nacional de Ciencias Exactas, Físicas y Naturales.

Fugelsang, J. and Dunbar, K. (2005), 'Brain based mechanisms underlying complex causal thinking', *Neuropsychologia,* 43, 8, 1204-1213.

Goodman, N. (1983), *Fact, fiction and forecast,* Cambridge, MA: Harvard University Press.

Llinás, R. (2002), *I of the vortex: from neuron to self,* Cambridge, MA: MIT Press.

Papert, S. (1980), *Mindstorms: Children, computers and powerful ideas*, Cambridge, MA: MIT Press.

Papert, S. (1993), *Children's machine: Rethinking school in the age of the computer*, New York: Basic Books.

Piaget, J. (1949), *Introduction à l'épistémologie génétique*, (3 vol.), Paris: Presses Universitaires de France.

Popper, K. (1959), *The logic of scientific discovery*, Hutchinson, London.

Pontifical Academy of Sciences (2003), *Papal Addresses to the Pontifical Academy of Sciences 1917-2002 and to the Pontifical Academy of Social Sciences 1994-2002. Benedict XV, Pius XI, Pius XII, John XXIII, Paul VI and John Paul II* (2003), The Pontifical Academy of Sciences, *Scripta Varia*, 100, Vatican City.

Minsky, M. (1986), *The society of mind*, New York: Simon & Shuster.

Negroponte, N. (1995), *Being digital*, New York: Knopf.

Sheridan, K., Zinchenko, E. & Gardner, H. (2005), 'Neuroethics in education', in: J. Illis (ed.) *Neuroethics in the 21st century. Defining the issue in theory, practice and policy*, Oxford: Oxford University Press.

Strauss, S. (2005), 'Teaching as a natural cognitive ability: Implications for classroom practice and teacher education', in: D. Pillemer and S. White (eds.), *Developmental psychology and social change* (pp. 368-388), New York: Cambridge University Press.

Session V

RESEARCH PROCEDURES: THEORIES AND THEIR VERIFICATION, SERENDIPITY

ON THE UNPREDICTABILITY OF INDIVIDUAL RESEARCH

MICHAEL SELA

I would like to distinguish between predictability by science and predictability of science. In the first case science is used to predict, e.g. climate, tsunamis or earthquakes.

In the second case, we wonder what direction science will take.

Undoubtedly, the great discoveries of science lead, predictably, to research trends resulting from them. Thus, the discovery of the double helix led to thousands of studies, ultimately leading to the breaking of the genetic code, and – in turn – after close to fifty years, to the elucidation of the human genome. Similarly, in physics, after the discovery of the theory of relativity or the theory of atoms, it was to be expected that – usually only after several years – a stream of studies resulting from these theories – would appear in a predictable fashion. The same is true of exciting new techniques. In life sciences discoveries like the cell-sorter (FACS) or the polymerase chain reaction (PCR), to give just two examples, have revolutionized many areas of experimental research. I would like to generalize these observations by stating that – at a 'macro' level – it is safe to assume that science is, to a large extent, predictable.

My contention is that this is not true at the 'micro', the individual research level. As most scientists are expected to write grant proposals – in which they describe their plans for research and the results they expect to reach – it is of interest to ascertain to what extent their predictions resemble the actual results. It would be depressingly boring if there would be too much resemblance between the plans and the subsequent reality. In all fairness, it must be stated that 'predictable' is not necessarily 'predicted', and if the results are actually opposite to what was predicted, in many cases this leads to breakthroughs of uncommon interest. We must be continuously watchful because very often the 'unpredictable' is lost because of lack of attention. The discovery of Fleming in 1928 of penicillin is due to his having paid attention to a Petrie dish with transparent areas in which the bacteria disappeared. This stresses the importance of serendipity which I define as 'luck meeting the prepared mind'.

I would like to give a few examples from my own research experience, and I refer to the discovery of the first synthetic polypeptide antigens, to the discovery of determinant-specific genetic control of immune response, to the discovery of a synthetic copolymer of amino acids that became an efficient drug against the exacerbating-remitting stage of multiple sclerosis, and to the discovery of a synergistic effect of a specific monoclonal antibody and of a chemotherapeutic drug in fighting cancer.

Synthetic Polypeptide Antigens

The purpose of the study was to make a protein a better antigen. It was then that we promoted the notion of immunogen and immunogenicity. We wanted to increase the immunogenicity of a protein, and we chose gelatin, a very poor immunogen, to which we attached chains of polytyrosine [1]. A limited polytyrosylation converted gelatin into a potent immunogen which provoked in experimental animals the formation of gelatin-specific antibodies. A more intensive polytyrosylation led to an immunogen which led solely to anti-tyrosine peptide antibodies [1,2]. The inevitable conclusion was that gelatin could be replaced with a synthetic branched polyamino acid and when we attached peptides including tyrosine to such a polymer, we obtained a synthetic branched macromolecule which was a potent and specific immunogen in several animal species [3]. So, in this case we wanted to improve the antigenicity of proteins, and we ended up with a whole array of synthetic antigens, which permitted us to elucidate many molecular aspects of antigenicity [2,4,5].

We could learn a lot about the role of size, composition, and shape, as well as about the accessibility of those parts of the molecule crucial for immunogenicity. As a matter of fact, we learned that it was possible (provided one was prepared to invest the necessary effort) to prepare synthetic immunogens leading to antibodies of essentially any specificity.

Although in most cases a good immunogen had a molecular mass of at least several thousand Daltons, dinitrophenyl-hexalysine and arsanil-trityrosine were by themselves capable of triggering an efficient immune response. The minimal size for a molecule to be immunogenic depends, therefore, largely on its chemical nature.

Although electrical charge may be important in defining the antigenic specificity of an epitope, charge is not a minimum necessary cause for immunogenicity; we could prepare water-soluble amino acid copolymers devoid of charge that were immunogenic. Polymers of D-amino acids were immunogenic only when they were administered in minute amounts and they led to no secondary response.

In the early days there was a wonderful feeling working on synthetic antigens because practically nobody else was working on the subject, but later on it was as pleasant and satisfying to know that so many laboratories had become interested in the synthetic approach to immunological phenomena. One of the most fascinating aspects of our studies with synthetic antigens had to do with the steric conformation of the immunogen and of its epitopes. We distinguished between conformational (conformation-dependent) and sequential determinants [6] and showed how the same peptide (Tyr-Ala-Glu) may lead to antibodies recognizing the sequence (when attached to multichain poly-DL-alanine) or recognizing an epitope defined by conformation (when the tripeptide was polymerized to give an α-helical structure). In addition, we could demonstrate for the first time, by circular dichroism, how antibodies to α helical polymer could help transconform into a helical shape a small polymer that was not yet helical [7]. These studies led us directly to study proteins and to synthesize a macromolecule in which a synthetic 'loop' peptide derived from hen egg white lysozyme was attached to branched polyalanine [8]. The resulting antibodies reacted with intact lysozyme through the 'loop' region, but the reaction was totally abolished when the disulfide bond within the 'loop' was opened, and thus the three-dimensional structure was collapsed.

Genetic Control of Immune Response

Even though some hints could be found in earlier literature, the actual establishment of the genetic control of the immune response became possible only through the study of synthetic antigens, simple chemically, in inbred strains of mice and guinea pigs, simple genetically.

I would now like to tell the story how it all started. In the summer of 1961, when I returned to the Weizmann Institute in Israel, from a year spent at the National Institutes of Health in Bethesda, Maryland, I stopped in London to discuss with John Humphrey and Brigitte Askonas a collaborative effort to follow the fate of strongly radioactive synthetic polypeptide antigens and to find out whether antigen molecules must be present in antibody-producing cells. Ultimately, this project was brought to a successful fruition [9), but in its initial stages, Hugh McDevitt, who joined Humphrey from Boston, injected cold poly(Tyr,Glu)-poly(DL-Ala)-polyLys, (T,G)-A--L, into rabbits to study their immune response. Several weeks later, Humphrey informed me at a WHO meeting in Geneva that the sandylop rabbits they used did not produce antibodies, and we considered the genetic makeup of the animal as one possibility to explain this result. Within a short time, it was clear that New Zealand rabbits produced as many antibodies as did our rabbits in Rehovot, and Himalayan rabbits were almost an order of magnitude better. At this moment, it was natural for McDevitt to switch to inbred strains of mice.

In our studies [10, 11] we first showed determinant-specific (antigen-specific) genetic control of immune responses by making use of multichain polyamino acids as antigens and inbred mice as experimental animals. (The first paper became a Citation Classic, Curr Cont. 1987). The multichain synthetic polypeptides we investigated, possessed at the tips of their polymeric side chains, small amounts of tyrosine, histidine, or phenylalanine. These antigens were denoted (T,G)-A--L, (H,G)-A--L, and (Phe,G)A--L. We noted that when histidine was substituted for tyrosine, genetic control was completely reversed, whereas replacement with phenylalanine led to a material strongly immunogenic in both the strains investigated.

Some time later, Hugh McDevitt, using these multichain polypeptides, was able to show for the first time the link between the immune response and the major histocompatibility locus of the mouse, which in turn led to our present-day understanding of immune response genes and their products. Of all the contributions of synthetic polypeptides toward our present-day understanding of immunology, none has been more important than the discovery and the definition of the genetic control of the immune response, which in turn was a crucial trigger toward a better understanding of the cellular basis of immunological responsiveness.

So, the initial project was to find out whether a cell producing antibodies has some antigen in it, and we ended up with discovering the genetically defined differences in the immune response.

Drug Against Multiple Sclerosis

We tried to build synthetic amino acid copolymers that would resemble myelin basic protein (MBP) and would induce, similarly to this protein, experimental allergic encephalomyelitis in animals (EAE), and only after we failed, we realized that they cannot initiate the disease, but they can suppress it.

In our early studies, of special interest was the immune response to lipid components, which was not easy to either elicit or investigate because of solubility problems. However, conjugates in which synthetic lipid compounds were attached onto synthetic copolymers of amino acids elicited a specific response to lipids such as cytolipin H, which is a tumor-associated glycolipid [12], or sphingomyelin. Furthermore, we demonstrated that both the sugar and lipid components of such molecules contributed to their specificity. The resultant anti-lipid antibodies were capable of detecting the corresponding lipids both in water-soluble systems and in their physiological milieu. This was fascinating because it gave us a glimpse into some disorders involving lipid-containing tissue and consequently led to our interest in demyelinating diseases, namely, disorders in which the myelin sheath, which constitutes the lipid-rich coating of all axons, is damaged, resulting in various neurological dysfunctions. We thus thought that EAE, caused by MBP might actually be induced by a demyelinating lipid and that the positively charged MBP might serve only as a schlepper (carrier) for an acidic lipid (e.g. phospholipids). We prepared several positively charged copolymers of amino acids and tested whether we could induce EAE when the copolymers were administered into experimental animals (guinea pigs and rabbits) in complete Freund's adjuvant, similarly to the successful administration of MBP, but we failed. On the other hand, the injection of several positively charged amino acid copolymers in aqueous solution into mice, rabbits, and guinea pigs resulted in efficient suppression of the onset of the disease, experimental allergic encephalomyelitis [13,14]. Later on, we could suppress the actual disease in rhesus monkeys and baboons. The copolymer 1 that was primarily used, denoted Cop 1, now called glatiramer acetate, and by industry 'Copaxone', is composed of a small amount of glutamic acid, a much larger amount of lysine, some tyrosine, and a major share of alanine. To our pleasant surprise, there is a significant immunological cross-reaction (both at the antibody level [15] and at the T cell level [16]) between Cop 1 and myelin basic protein. Interestingly, when an analog of Cop 1 made from D-amino acids was tested, it had no suppressing capacity nor did it cross-react immunologically with the basic protein. Cop 1 is not generally immunosuppressive; it is not toxic; actually it is not helpful in any other autoimmune disease except in multiple sclerosis and its animal model, experimental allergic encephalomyelitis.

The clinical trials with Cop 1 have included two preliminary open trials and two double-blind II trials, one involving exacerbating-remitting patients

[17] and another one in chronic progressive patients [18]. The results of the phase II trial in exacerbating-remitting patients demonstrated a remarkable decrease in the number of relapses and in the rate of progression in Cop 1-treated patients compared with the placebo control. Cop 1 is a promising low risk multiple sclerosis-specific drug for treatment of the relapsing disease. As an antigen-specific intervention, Cop 1 has the advantage of reduced probability of long term damage to the immune system.

After a successful, pivotal multicenter phase III clinical trial conducted in 11 medical centers in the United States [19], Cop 1 was approved by the United States Food and Drug Administration as a drug for multiple sclerosis. This was a moment of gratification and deep emotion for my colleagues and myself, as well as for our industrial partners, Teva Pharmaceutical Industries.

An important step in our understanding of the mode of action of Cop 1 was the observation that copolymer 1 induces T cells of the T helper type 2 that cross-react with myelin basic protein and suppress experimental autoimmune encephalomyelitis [20]. This was corroborated by clinical studies in multiple sclerosis patients [21]. It was of interest to observe that Th2 suppressor lines and clones induced by Copolymer 1 cross-reacted at the level of Th2 cytokine secretion with myelin basic protein but not with other myelin antigens [22]. This bystander suppression may explain the therapeutic effect of Cop 1 in EAE and multiple sclerosis (MS).

Cop 1 binds promiscuously to many different cells regardless of their DR restriction. It binds avidly and fast and can also displace already bound antigens, and this holds for all the myelin antigens that may be involved in MS; and yet, Cop 1 exerts its activity in an antigen-specific manner (it is not a general immunosuppressive agent and does not affect other experimental autoimmune diseases). Its specificity must, therefore, be envisaged in the context of the trimolecular complex MHC-Ag-T-cell receptor ('the immunological synapse'), namely, as interference with the presentation of the encephalitogenic antigen to the T-cell receptor, which is a specific interaction.

I recently summarized the story of specific vaccines against autoimmune diseases [23], as well as the successful use of Cop 1 (glatiramer acetate, Copaxone) in the treatment of multiple sclerosis for exacerbating-remitting patients [24]. The majority of the patients in the great clinical trial continue to be followed in an organized fashion for more than 7 years. Their risk of an MS relapse was over 1.5 per year at onset and is now less than 1 every 6 years. On an average, these patients have experienced no increase in neurological disability, whereas natural history profiles would have predicted substantial worsening. The accumulated experience with glatiramer acetate (Cop 1) indicates that its efficiency is apparently increased as a function of usage time, while the favorable side effect profile is sustained.

Personally, the whole odyssey of Cop 1 and its use in MS has been a source of great satisfaction and emotion. The awareness that over one hundred thousand MS patients feel better because of a drug/vaccine that we conceived and developed, moves me deeply. Twenty-eight years have passed from the moment of the idea to the approval of Cop 1 by the Food and Drug Administration. I have a feeling that discoveries resulting from basic research take a longer time to fruition, but on the other hand, they are probably more original in terms of concept.

Synergy Between a Monoclonal Antibody and Chemotherapeutic Drugs

In this case we covalently bound for long period chemotherapeutic drugs to anti-tumor antibodies, using the latter mainly as missiles to target the drug. Over the years we found out that some monoclonal antibodies are very efficient as anti-cancer drugs, but the greatest effect was obtained when we used the combination of the antibody and the chemotherapeutic drug [25].

The idea of binding anti-cancer therapeutic drugs covalently to antibodies reacting with cancerous cells has appealed to me from an early time. Instead of having the drugs given systemically, spread throughout the whole body, immunotargeting would focus the supply of the drug exclusively to the cancer area. However, we did not get to immunotargeting until many years later, when we bound daunomycin and adriamycin via a dextran bridge to antibodies against antigens of leukemia, lymphoma, and plasmacytoma cells. We showed that these are effective as 'guided missiles' both in vitro and in vivo [26].

Later on we moved to monoclonal antibody against the extracellular domain of the epidermal growth factor receptor, denoted today ErbB1, and found that its conjugate with daunomycin was quite efficient but so was the antibody by itself [25]. A strong synergistic effect was observed when the anti-ErbB1 antibodies were administered together with cis-platin. This observation became of great interest because of its therapeutic potential (e.g. in the review article by Mendelsohn and Baselga [27]). Over the years, I became more and more *convinced* that what matters most is the nature of monoclonal antibodies.

Conclusions

It is of crucial importance to have well defined plans in research, but it is at least as important to be flexible and open minded, and to conduct research in a way that leads to optimal results. These often lead to unexpected discoveries, as I hope I have showed in the four examples I have illustrated here.

REFERENCES

1. Sela, M., and Arnon, R., *Biochem. J.*, 75, 91 (1960).
2. Sela, M., *J. Biol. Chem.,* 278, 48507 (2003).
3. Sela, M., Fuchs, S., and Arnon, R., *Biochem. J.,* 85, 223 (1962).
4. Sela, M., *Adv. Immunol.*, 5, 29 (1966).
5. Sela, M., *Science,* 166, 1365 (1969).
6. Sela, M., Schechter, B., Schechter, I., and Borek, F., *Cold Spring Harbor Symp.* Quant. Biol., 32, 537 (1967).
7. Schechter, B., Conway-Jacobs, A., and Sela, M., *Eur. J. Biochem.*, 20, 321 (1971).
8. Arnon, R., Maron, E., Sela, M., and Anfinsen, C.B., *Proc. Natl. Acad. Sci. USA,* 68, 1450 (1971).
9. McDevitt, H.O., Askonas, B.E., Humphrey, J.H., Schechter, I., and Sela, M., *Immunology*, 11, 337 (1966).
10. McDevitt, H.O., and Sela, M., *J. Exp. Med.,* 122, 517 (1965).
11. McDevitt, H.O., and Sela, M., *J. Exp. Med.*, 126, 969 (1967).
12. Arnon, R., Sela, M., Rachaman, E.S., and Shapiro, D., *Eur. J. Biochem.,* 2, 79 (1967).
13. Sela, M., Arnon, R., and Teitelbaum, D., *Bull. Inst. Pasteur*, 88, 303 (1990).
14. Teitelbaum, D., Meshorer, A., Hirshfeld, T., Arnon, R., and Sela, M., *Eur. J. Immunol.,* 1, 242 (1971).
15. Teitelbaum, D., Aharoni, R., Arnon, R., and Sela, M., *Proc. Natl. Acad. Sci. USA*, 85, 9724 (1988).
16. Teitelbaum, D., Milo, R., Arnon, R., and Sela, M., *Proc. Natl. Acad. Sci. USA,* 89, 137 (1992).
17. Bornstein, M.B., Miller, A., Slagle, S., Weitzmann, M., Crystal, H., Dexler, E., Keilson, M., Merriam, A., Wassertheil-Smoller, S., Spada, V., Wein, W., Arnon, R., Jacobsohn, I., Teitelbaum, D., and Sela, M., *N. Engl. J. Med*. 37,408 (1987).

18. Bornstein, M. B., Miller, A., Slagle, S., Weitzmann, M., Drexler, E., Keilson, M., Spada, V., Wein, W., Appel, S., Rolak, L., Harati, Y., Brown, S., Arnon, R., Jacobsohn, I., Teitelbaum, D., and Sela, M. *Neurology,* 41, 533 (1991).
19. Johnson, K.P., Brooks, B.R., Cohen, J.A., Ford, C.C., Goldstein, J., Lisak, R.P., Myers, L.W., Panitch, H.S., Rose, J.W., Schiffer, R.B., Vollner, T., Weiner L.P., Wolinky, J.S., and the Copolymer 1 MS Study Group, *Neurology,* 45, 1268 (1995).
20. Aharoni, R., Teitelbaum, D., Sela, M., and Arnon R., *Proc. Natl. Acad. Sci. USA.,* 94, 10821 (1997).
21. Neuhaus, O., Farina, C., Yassouridis, A., Wiendl, H., Bergh, F.T., Dose, T., Wekerle, H., and Hohlfeld, R., *Proc. Natl. Acad. Sci. USA,* 97, 7452 (2000).
22. Aharoni, R., Teitelbaum, D., Sela, M., and Arnon, R., *J. Neuroimmunol.*, 91, 135 (1998).
23. Sela, M., *C.R. Acad. Sci. Paris Life Sci.* 322, 933 (1999).
24. Sela, M., and Teitelbaum, D., *Expert Opin. Pharmacother.,* 2, 1149 (2001).
25. Aboud-Pirak, E., Hurwitz, E., Pirak, M.E., Bellot, F., Schlessinger, J., and Sela, M., *J. Natl. Cancer Inst.,* 80, 1605 (1988).
26. Levy, R., Hurwitz, E., Maron, R., Arnon, R. and Sela M. and Sela, M., *Cancer Res.* 35, 1182 (12975).
27. Mendelsohn, J., and Baselga, J., *Oncogene,* 19, 6550 (2000).

SCIENCE AS PREDICTION AND THE UNPREDICTABILITY OF SCIENCE

STANLEY L. JAKI

This paper is about two different features of science. The difference is tied to two very different meanings of the same word, 'humanly', when a prediction is considered to be humanly impossible. One meaning relates to that most eminently human quality which is to act freely. And since such acts cannot be measured, they are not to be considered in reference to scientific predictions. Most cultivators of psychology would disagree, but they do so by changing the meaning of 'free' in reference to human acts. Yet unless they make that change freely, the change loses its meaning because it implies that all human minds, including those of psychologists, are acting like so many machines, blindly and inevitably. This facet was not noticed by William James, who a hundred years ago reported about psychologists (he was a chief among them) that they almost to a man held that new meaning of human freedom, which rendered the free will meaningless. The coming of Freudian and other forms of psychoanalysis did not shore up the case of human free will. Then in the 1930s there came a brief dallying among physicists with the idea that quantum mechanics allowed the free will to operate within a very narrow limit set by Heisenberg's uncertainty principle. Eddington, who first used that principle in that sense, took the view, and within one year, that the effort made no sense.

Another meaning of 'humanly' is purely pragmatic. There was a time when it was practically impossible for humans to predict lunar or solar eclipses. Ptolemaic astronomy removed that impossibility, but it did not thereby perform a prediction properly speaking. The Ptolemaic system of epicycles and eccentrics implied that there should follow eclipses at a given place and time in the sky. The prediction was so accurate that Copernicus' theory did not improve on it.

Until about a hundred years ago it did not seem possible to predict by science that the perihelion of Mercury might not be re-entrant. Although Newton spoke of that possibility, he also stated that a new physics was needed to predict it, but of that new form he had no inkling. Efforts to predict that advance by modifying the inverse square law, however slightly, led to impossible consequences in other respects. Strictly speaking even general relativity did not predict the advance of the perihelion of Mercury. The theory merely presented that advance as being implied in the essential parameters of the theory. It is another matter that what is implied in a theory is not immediately noticed by all physicists. Those who notice something which others have not yet noted deserve credit, but not for having predicted something that was strictly unpredictable, such as events contingent on free human actions.

Scientific predictions are better called extrapolations. Illustrations of this are all great advances of Newtonian physics, usually called predictions. They are just extrapolations, though with one difference. They are *exact* numerically. Examples are Euler's work on the libration of the moon, and Le Verrier's claim that another planet, still unknown in his time, causes the perturbations in the motion of Uranus. The claim had to include the determination of the spot and time where one was to look for a body which, once spotted, was named Neptune. The inability of classical physics to measure, which is an operation with numerically exact data or parameters, the influence of the ether on the earth's motion led to the abandoning of Newtonian physics. The latter implied exactly the opposite of what Michelson established with his interferometer. There is some exactness even in the so-called null result. The ability of Balmer's formula to indicate some spectral lines *exactly* led to Bohr's formulation of his atom model.

Exact science is 'predictive' though not in that strict sense in which, say a prophet, who is truly a prophet, predicts something. A true prophet has to deal with future events contingent on free human acts. A memorable example of this is Jesus' prediction of the destruction of Jerusalem and of the Temple. Humanly speaking it should not have happened, except for some human acts that are recorded in Josephus Flavius' *Jewish Wars*. I am not talking of the sloppiness of the guards of the Southern Gate, who left their post because of a torrential downpour in the middle of the night. This gave opportunity to a large number of ruffians to enter the City and reinforce there similar elements. Far more specifically human, that is strictly unpredictable, were some actions of John of Gischala, the general of Jewish forces in Galilee. After having been roundly defeated by the Romans, he somehow escaped their grip, fled to Jerusalem where he told the leaders that he had just inflicted a crushing defeat on the Romans. This lie emboldened the defenders of the City to reject peace offers from Titus, who to the end wanted to save the Temple. He could not, however, prevent a soldier from throwing a burning piece of timber into the Temple, which then went up in flames.

In terms of a science that claims omniscience about all physical events, big or small, it may not have been absolutely impossible to predict that Peter would find a didrachma in the mouth of the first fish he would catch on Jesus' command. Even stranger things can be found in the mouths of fish. But if that ring had been thrown into the water by a free human act, then no science, however perfect, could have foretold Peter's success. So much for prophecy as distinct from mere prediction, which it is the business of science to make, though only of a specific form of science, exact science. Science is exact insofar as it relies on exact measurements, which are exact because they are operations with numbers.

Other so called sciences, such as political science, represent a very poor class by comparison. It was surely a poor prediction when Henry Kissinger predicted in 1988, two years before the collapse of the Soviet Union, that for the next hundred years the Soviets would remain the *other* superpower. Kissinger had been for years professor of political science at Harvard before serving as Secretary of State.

Exact science is predictive through the use of the only form of words, 'numbers', that imply a quantitative form of exactness, which represents a very special meaning of the word *exact*. Only of numbers is it not possible to state the phrase, 'more or less'. This point was made more than two thousand years ago in Aristotle's *Categories,* though he himself did not perceive how important a point he had made. He surely proved that human minds do not work like machines. Had he noted the importance of his remark, intellectual history might have become very different and exact physical science might have emerged two thousand years earlier than it did. Let it not be forgotten that Archimedes' method of computing the volume of a cone by approximation as if a cone were a set of slim disks all gradually smaller in diameter, came very close to the method of going to the limit which is the basis of infinitesimal calculus.

The ability of exact science to predict, that is, to unfold, should seem so obvious that it would be a waste of time to dwell on it. Quite different is the case with the other part of the title of this lecture, namely, the unpredictability of science. The unpredictability of science prior to the 17th century was part of a presentation I gave nine years ago in this Academy, under the title, ‘The Earth-Moon System and the Rise of Scientific Intelligence’. The paper dealt in part with the impossibility of predicting the accretion of a moon to the earth in such a way as to give the earth that very moon which orbits around it in the way it does. In spite of the great increase of the number of moons that have been for the last three decades spotted around other planets, the parameters of the earth-moon system remain unmatched. Yet the rise of Ptolemaic astronomy, which is the basis of all exact science, would not have been possible without that earth-moon system, or rather without the special lunar and solar eclipses it implies.

Further, there was the need for geniuses, such as Eratosthenes and Aristarchus of Samos. Their emergence on the Greek intellectual scene gives the lie to the Baconian method of doing science. That method implies that if one looks long and hard enough, one would find. This logic is no better than the saying that even a blind chick would find a grain now and then. In another form the Baconian method implies that the chances of finding are the greater the larger is the number of those who are looking. Well, the very large number of those who lent their computer time to make sure that messages from extraterrestrials would not be missed, left our isolation intact. No serious historian of Greek science ever claimed that the work of Aristarchus of Samos could be foreseen by earlier Greeks. No serious historian of Greek science ever suggested that three hundred years after Aristarchus of Samos, there would be a Ptolemy who would synthesize the work of earlier Greek astronomers. Different may be the case in reference to the rise of Euclidean geometry. Unfortunately we know only the names of the two hundred or so Greek geometers who had flourished before Euclid.

Some historians of science claimed that Buridan’s epoch-making formulation of inertial motion in 1332 would have come anyhow. But such historians always showed a dislike for the epoch known as the Middle Ages, insofar as it was steeped in Christian, that is, Christ-centered faith. This faith, which was directly connected to Buridan’s achievement, was not predictable, though it was prophesied. Nor was the coming of Buridan predictable. If the Black Plague, which claimed one-third of Paris, had come not in 1347 but twenty years earlier, Buridan might have died before obtaining the chair of natural philosophy at the Sorbonne. Kepler’s three laws might never have become known to Newton, had Jeremiah Horrocks died not at the age of twenty-two but at the age of twenty. Young Newton might have died at the age of nineteen when he escaped from plague-stricken Cambridge to his father’s farm. Then an apple might not have fallen on his head to prompt him to think about gravitation. Neither the coming of geniuses, nor the fall of a given apple at a given place and time is predictable, if the context implies the free decision of one to take his seat under a given apple tree.

Once the three laws of motion were in place in Newton's *Principia*, the progress of science became a logical unfolding of the implications of those three laws. The steps of that unfolding, which had been taking place at an accelerated rate as time went on, was not something foreseeable in Newton's time. For decades the leading French geometers refused to accept Newton's *Principia* as a book on physics. No predictions could be made about the discoveries to be made in terms of Newton's physics, although some were increasingly looked for. Suffice it to recall the discovery that the mutual perturbation of planets cannot go beyond a maximum and that therefore the solar system was stable. This discovery, which Newton still claimed to be impossible to make, was done first by Lambert, and perhaps independently of him by Laplace. This detail would lead to the subject of simultaneous or multiple discoveries, a subject that would provide further doubts about the predictability of science.

No reputable historian of electromagnetism has claimed that following Weber, who among other things devised an electro-magnetic telegraph, Maxwell would formulate his electro-magnetic theory, although from Weber's work it followed that the speed of light was what was predicted by Maxwell's theory. Maxwell's theory, or rather his equations have not lost their lustre of originality and beauty. Einstein himself was so overawed by their perfection as to hold that they must retain their original form in all reference systems. Relativity theory has other motivations than purely scientific. Einstein himself did not realize early enough that his theory should have been called the theory of invariance. Even geniuses may fall short of their potentials, which is hardly a subject for predictions. Einstein himself admitted that special relativity would have been formulated even without him. The evidence of this is in Whittaker's *History of the Theories of the Ether and Electricity*. It is another matter whether one should give too much credit to Einstein's claim that the general theory needed him and him alone. It may very well be that Hermann Weyl would have formulated it, had Einstein not done it early enough.

No one who works in any branch of exact science can predict what happens there within the next fifty years except for some trivial points. A memorable registering of this came in 1950 when the editors of *Scientific American* asked eight prominent scientists to summarize the main advances made within their respective fields during the first half of the twentieth century. About astronomy the Harvard astronomer Harlow Shapley wrote:

> Scarcely a question is asked of a doctoral candidate today [i.e. 1950] that would have made sense to the giants [of astronomy] of 1900. They would have been baffled, helpless and perhaps suspicious in the face of inquiries concerning photomultipliers, quantum theory, solar spicules, the carbon cycle, shell stars, the expanding universe, radio 'hot spots', the Schmidt reflector, Pluto, cosmic rays, and other common topics. Pride in our advances should be mellowed, by the contemplation of how much beyond us the astronomical world of 2000 A.D. is likely to be.

In 1950 no giant of astronomy would have predicted black holes, the 2.73°K cosmic background radiation, orbital telescopes, space probes, semiconductors, nanotechnology and so forth. Today nobody knows for sure whether in 2050 more than a few historians will remember string theories, let alone speculations about multiple and parallel universes. As for string theories, one cannot help admiring the wizardry of those who manipulated scores of dimensions in such a way as to lead to the mathematical formalisms underlying the four known physical forces, gravitational, electromagnetic, weak, and nuclear. String theorists have not claimed so far to predict the constants implied in those forces. As to cosmological theories that imply the change of the values of those constants with time, one can only predict a regular, but not a chaotic change. Predictions must obey the laws of logic. A chaos theory that implies absolute randomness is a contradiction in terms. Also the prediction cannot be random, let alone chaotically so. This would bring up the absolute priority of rationality over irrationality, of which more later. Inadmissible should seem the use of the term 'universe', which does not stand for a strict totality. Cosmological theories of multiple universes should seem suspect on purely logical grounds. At any rate, nothing specific has been so far predicted about any of those other universes.

To give a facetious touch to all such reflections by a historian of science on scientific predictability, I would like to recall that a hundred years ago Marcelin Berthelot, a notable French chemist in his day, but largely forgotten today, stated that around the year 2000 all human food would be in the form of pills. I am glad that the food service of the Vatican does not try to prove that prediction. Around 1900 Samuel Newcomb, the leading American astronomer at that time, claimed that machines heavier than the air could never fly. Fifty years later Vannevar Bush of MIT and presidential advisor on science, held that intercontinental ballistic missiles were impossible to construct. Twelve years later two other presidential advisors, Wiener and Teller, gave two diametrically opposite views on space defense. President Kennedy quipped: 'I am therefore free to do whatever I want'. In 1959 Robert Leighton of Caltech and author of a highly regarded textbook on advanced physics, claimed that the study of fundamental particles had essentially been completed. He had no idea of the coming of giant accelerators, such as the one in Batavia, Illinois, or CERN in Geneva, built precisely to detect ever more elusive new particles of which it was well stated that none of them were really fundamental.

What alone can safely be predicted about the future form of physics is that it will become more and more complex mathematically. To say, as Roger Penrose did, that the future form of physics will be a new form of quantum mechanics which would include general relativity leaves intact the apparent irreducibility of summation and integration to one another. There is no middle ground between discontinuity and continuity. And as long as Gödel's incompleteness theorem remains valid, one can safely predict that a form of physics which would be necessarily true cannot be achieved. In other words, theoretical physics remains an open-ended venture. This conclusion was reached two years ago by Stephen Hawking after he had bemoaned half a year earlier the end of physics, because of Gödel's theorem. It means the end of only that form of physics which has on it the mark of hubris.

One can safely predict that Prof. Hawking will not see the realization of his prophecy, made in Beijing a few months ago, that he would be condemned by the Vatican in the same way as was Galileo. To begin with, the Vatican did not forbid Prof. Hawking to probe into the physics of nothing. John Paul II merely said that the creation out of nothing, insofar as that nothing is really nothing, cannot be handled by physics. Such is a most sensible statement, but beyond the common sense of some physicists, Guth for instance at MIT, who boldly speak about their power to create entire universes literally out of nothing. Once the talk of the town, but now largely forgotten, the steady state theorists claimed that their theory indicated the emergence of hydrogen atoms literally out of nothing. They did not like the comment of Pius XII on the theory as being gratuitous, although it could be viewed as plainly irrational. It is to the eternal credit of Benedict XVI that from the *aula magna* of the University of Regensburg, he reminded a Western world wallowing in all sorts of irrationalities – pragmatism, logical positivism, deconstructionism and the like – that rationality grounds all human discourse, whether about religion or anything else, including irrationality, whether science coated or any other kind.

Going from strictly exact forms of science to their partly exact forms, such as Darwinian theory, it is worth noting that it fails to predict future forms of species, let alone future genera and higher classes. This is all the more interesting because the scientific character of Darwin's theory lies in a fact, not fully recognized by Darwin himself. The fact is that there is a quantitatively measurable difference between parents and offspring, and that the impact of the physical environment, measurable in principle, must be different on the offspring and on the parent. Darwinian theory is science, but an exact science only in a very narrow and limited sense, a point which infuriates most Darwinists.

The inability of exact science to predict its own future, while it can predict far away events, in space and in time, is one of its serious shortcomings. About those shortcomings Polykarp Kusch, who received the Nobel Prize in 1955, said in 1963 that the power and impotence of physics are its two main sides and that both sides deserve equal scrutiny. He could have said that the future of mankind is as much in the hands of exact science as it is not. There is no scientific futurology. This is why the study of history is not a science though it can be a form of reasoned discourse. Rather unreasonable should seem long-range projections about physical situations, such as global temperature changes, that depend on an enormous variety of factors, some well known, some only guessed, and still others wholly unknown at a given moment.

Nothing can be predicted about another and very different kind of warming. It is the increasingly overheated availability of tools and gadgets delivered by science. It took a hundred years to put a billion telephones into human hands. It took only ten years, the last ten, to market another billion telephones. Similar figures, all indicative of an exponential rise, could be provided with reference to TV sets, automobiles, cameras and so forth. Libraries that have taken hundreds of years to develop now can be matched by electronic means within a few years. The text of all books printed in five hundred years may be available in five years if Google has its way. A single laptop can now house entire libraries.

It seems that the *génie* is out of the bottle and that nobody knows how to put it back there. Exactly seventy years ago when quantum mechanism was born, a strange proposal was made about that *génie*, although at that time nobody could foresee all that technological cornucopia which quantum mechanics would dump on mankind. In that year, we are in 1927, the British Association for the Advancement of Science held its annual meeting in Ripon, a quaint provincial town in Yorkshire. It was a custom that on

Sunday the Association would attend a church service, which in this case was led by the Anglican bishop there. He suggested that for three years all laboratories be closed so that scientists may have enough time to think over what they were doing. The next day it was the turn of Oliver Lodge, the grand old man of British science at that time, to speak at the banquet. He rejected the bishop's suggestion on the ground that it was impossible to halt discoveries. He was right, but I wonder whether he suspected that seventy years later scientific innovations would increase exponentially.

Science cannot do anything about this explosiveness. And I have not touched on what is being done in terms of genetics. Are we going to be duplicated, just because it is possible to do it? (Such was the argument on behalf of cloning of a member of President Bush's commission on the subject). This would not merely duplicate our problems, but make them grow hundredfold. Scientists cannot do much about a much simpler problem: namely, how to persuade science reporters to do their work responsibly.

In these times when so much is said about the interaction of exact science and the humanities, it is well to ponder a most human aspect of science. For the same reason that we humans cannot predict our own future, we cannot predict the future of our wonderfully exact science. This is so because it is a science made by humans who do not carry out everything exactly, not even their exact science. For better or for worse, scientists, as well as science, come with a built-in unpredictability.

Session VI

PUBLIC PERCEPTION AND POLICY IN THE CONTEXT OF UNCERTAINTY

SOME ISSUES RELATING TO PREDICTABILITY AND CERTAINTY IN SCIENCE

M.G.K. MENON

Science: The Triumph of the Reductionist Approach

We are aware that the human brain is gifted with an intrinsic sense of curiosity. It is because of this that human beings have wondered about their surroundings beyond the needs of survival. Nature around appears at first sight to be highly complex: in colour, form, sound, motion and the like – and therefore likely to pose serious difficulties for analysis and understanding. Human beings, however, have also been gifted with the ability to reduce this complexity to manageable proportions – through the power of logic and being able to look at the essentials that characterize any complex situation. It is, thus, that we have the apocryphal story relating to Newton and the falling apple, from which he deduced the far-ranging law relating to gravity. The great early scientific discoveries (like the Archimedes Principle, Newton's Laws of Motion, Galileo's experiments in Mechanics, and many such others) were all characterized by this effort to draw over-arching conclusions of a wide-ranging nature, by looking at the problem in its essentials.

Experiments are thus carried out on the essential parts of the phenomenon sought to be understood and observations made, with measurements of quantities; necessary methods and tools for the purpose are developed and constantly improved to increase precision. Often the latter lead to completely unexpected observations, uncovering new phenomena.

Lord Kelvin had remarked:

> When you can measure what you are speaking about, and express it in numbers, you know something about it. But when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of

> knowledge, but you have scarcely, in your thoughts, advanced to the stage of science.

Efforts are also made to have large numbers of persons make the measurements and in many different ways; and to have a large number of measurements from which a statistical analysis can be made. It is only when there is agreement, amongst all of the data, that one feels confident or certain about the final outcome; and further, to ascribe a defined uncertainty to the measurement. An important part of any result thus obtained, is the error that can be ascribed to it, of a systematic or statistical nature.

Scientists have been able to make measurements of very different types over a range of fields such as astronomy, botany, chemistry, physics, zoology and the like. It was thus that the enormous wealth of nature could be separated into different categories; and within these, classified in appropriate ways which would enable discovery thereafter, of simple relationships that make it possible to make predictions.

Thus, Tycho Brahe devised new and accurate instruments like the quadrant and the sextant to make large scale astronomical measurements. From this, Johannes Kepler, after incredible and innumerable wrong guesses, gave up circles and ovals and tried the ellipse. Then, all of Brahe's work clicked into place. Kepler was able to generalize planetary motions. At this point, the laws only described the motions, but did not offer any explanations for their cause.

In a very different field, Mendeleev was able to take all the known available data concerning the various elements and classify them in a table in which there were a large number of gaps for new elements yet to be discovered but with broadly defined properties; this enabled prediction of new elements. This was the power of predictability based on meaningful classification.

Again, the famous collectors in the fields of botany and zoology, mounted expeditions to a variety of places in the world, to bring together great collections containing innumerable specimens and were able to classify these. Amongst the greatest of these was Charles Darwin, who was able to formulate the theory of evolution.

In parallel to this observational and experimental approach, there was the theoretical approach which was founded significantly in mathematics. Galileo, rightly regarded as founder of modern science, specifically referred to the use of mathematics in expressing experimental results. About the 'book of nature', he had remarked: 'It cannot be read until we have learnt the language and become familiar with the characters in which it is written. It is written in mathematical language'. As the concepts about the physical universe grew in complexity and subtlety, so did the mathematics to describe them. In the 18th and 19th centuries, there were gifted individuals such as Euler, Fourier, Gauss, Hamilton, Lagrange, Laplace and others who contributed to the development of physical sciences in the theoretical framework as also the mathematical underpinning of this. Later, when the quantum and relativity revolutions took place in the 20th century, the great mathematical edifice that had been developed independently purely within mathematics was made use (Poincare, Riemann, Minkowski and others).

It is this interplay between theory and experiment, underpinned by mathematics, that constitutes the essence of modern science. The extraordinary accuracies of predictions and observational verification of these made Dirac remark: 'This must be ascribed to some mathematical quality in nature – the quality, which a casual observer of nature would not suspect but which, nevertheless, plays an important role in nature's scheme of things'.

It is interesting to comment on the prediction and discovery of antimatter. The prediction was made by Dirac as a consequence of a purely theoretical approach representing the sheer power of intellect. The actual discovery of the existence of the positive electron was made by Anderson by adding a magnetic field to a cloud chamber and, thus, opening a new window through which he observed not only the positron, but also the particle of intermediate mass (between the electron and proton), now known as the muon. Again, the addition of counter control to the cloud chamber enabled Blackett and Occhiatini to demonstrate the creation of electron-positron pairs. Experimental discoveries of antimatter were, thus, entirely observational, and not as a result of a search for antimatter predicted by Dirac. In contrast, the search for the antiproton was planned with that objective.

It is possible to list out the very large number of cases of predictions based on mathematical and theoretical formulations as also through classification schemes (such as of Mendeleev and of elementary particles). One can also see the precision with which one can measure many physical entities such as the 'Lamb shift' or the 'g-2' of the electron and the muon (upto 0.6×10^{-9}) and, indeed of the precision with which one observes the black body microwave radiation characteristic of the Universe. All this illustrates the triumph of the reductionist approach in science.

Non-linear, Non-equilibrium Phenomena and Complexity

However, in spite of all this, and this is the edifice which has enabled science to be built up to where it is today, the real world is complex and I'll tell you a small story, relating to how I began to understand it that way.

Over 50 years ago, in 1951, I was sitting at a table in a canteen, in England, and I asked permission to sit because there was somebody else there. I never realized who that person was, though he happened to be one of the most famous persons I could have encountered then. We started talking and I was, of course, so full of what I was doing that I kept talking about it; until he asked me, 'Are you doing any teaching?', I said 'Yes, I've been asked to take a course in thermodynamics' and he asked me what I was teaching in thermodynamics. I explained the classical thermodynamics which one teaches at English undergraduate levels, concerning the first, second and third laws of thermodynamics, entropy and the like; and then he said, 'You know, young man, I think you have to realize that whilst all of what you are teaching is correct the real world is more complex than the standard laws of thermodynamics'. And then he gave me a little lecture for ten minutes on the whole area of complexity, non linear phenomena, and what now are referred to as dissipative systems. That person, whom I didn't recognize and know then at all, was Alan Turing. Alan Turing is a father figure in the field of computers, of automata, the Turing machine and so on; also famous for the work he did during the Second World War at Bletchley Park and the Enigma machine, breaking German military codes and the like. Unfortunately he committed suicide a little later. But the work that he was talking about, 'Turing bifurcations', and dissipative mechanisms, is what was referred to in later work of scientists like Prigogine. Turing had worked on these bifurcations in his classic paper of 1952 on morphogenesis.

What I want to say is that there is a great deal of what one would call simplicity, reductionism and direct predictability in science, but there is also the real world of complex systems which we encounter in real life, and that is what the public normally deals with.

The origin of this meeting goes back to some ideas proposed by the Academician Keilis Borok who had referred to similarities in a variety of complex systems and the underlying mathematics. He had given talks at the Academy on phenomena such as: earthquakes and seismicity; political, social and economic systems; patterns of crime; the behaviour of financial markets; traffic patterns; and the like. These are all highly complex situations which are also volatile in their behaviour. The question was how does science deal with such situations.

It has been pointed out at this meeting by Prof. Zichichi that there are different levels for understanding scientific phenomena. Thus, whilst we may make very detailed analysis of QCD, these relate to interactions involving quarks and gluons, but they do not come into the picture when we consider interactions between protons and neutrons. It is, thus, that one has the Newtonion-Galileian picture operating in normal mechanics but need an entirely new understanding through quantum mechanics when we go to the levels of atoms and electrons. One of the important questions to be considered is whether the complex phenomena encountered normally in society can be analyzed in reductionistic terms. Are there additional principles of nature that are superposed on complex systems which, of course, will follow all the other laws of nature that one has reduced through reductionist science? Is the whole greater than the sum of the parts?

In the earliest periods of human history, one has come across many examples of innovation and discoveries that have transformed human life and became the bedrock of civilization. These include 'the system of numbers', particularly 'the decimal place value system', and 'the zero'; 'the invention of the wheel'; 'the discovery of fire'; 'the selection of plants with economic value' and being able to grow them under controlled conditions; and the like. None of these developments are associated with the names of individuals. Since medieval times, and more particularly through the history of modern science, one has discoveries that are associated with individuals or schools of thought. But throughout this period, the discoveries remained largely within science and were of little concern to society. There are, no doubt, exceptions, particularly in relation to weapons of war which affected society. But, by and large, science was an internalized system.

Who will deny that the year 1666 was indeed The Year of Wonders (Annus Mirabilis) – for science. That was the year when Newton was responsible for spectacular scientific developments (Binomial Theorem, Fluxions/Differential Calculus, Theory of Colours; Integral Calculus/Inverse Fluxions; and Theory of Gravitation). All of these turned out to become the fundamental underpinnings of modern science. Yet none of these would ever be regarded by society then as being of importance to it. This was in contrast to the case of Galileo, who had just preceded Newton, who was an experimental scientist par excellence but was persecuted for his convictions which ran contrary to theological beliefs of that time. That, however, was a case involving those in authority and the hold they had on society at large. Similar to the case of Newton was that of Einstein who produced five spectacular papers that shook the foundations of physics in 1905. Yet these discoveries were of little concern to society at large.

This is in total contrast to the situation encountered today, where science cannot be regarded as so isolated. Scientific discoveries receive a great deal of publicity in the media (particularly the print media, radio and television). In addition, scientific discoveries very often lead to technological developments of a profound nature impacting on society.

With the development of technology through innovation, new artefacts continue to develop. Not all of these are of great use or desired by human society, at the time when they are developed. A large number would be acceptable because they represent improved versions of what was available earlier, or perform some new functions; for example, there are pharmaceutical products which have fewer side effects in terms of toxicity or are more effective or have a more desirable method of administration. There are other artefacts that are wholly new like: computers, TV sets, cell phones, internet and the world-wide web, genetic engineering and the like. These totally change the functioning and thought processes of society and are regarded as disruptive technologies.

It is these developments that often constitute the image of science in the public mind – an image of the extraordinary powers of the human mind to probe into the smallest and the largest, of living and non-living systems, to connect all of humanity, to imitate and mimic life-like processes, indeed leading to artificial life. With this image there is also, amongst many, a confidence in the invincibility of the technocratic fix and a certainty attributed to science.

Science and Social Problems

Cecil Powell, in an address that he gave in 1968 (Selected Papers of Cecil Frank Powell edited by Burhop, Lock and Menon, North Holland Publishing Company, 1972, p. 444), had remarked:

> Scientists ought also to play an increasing role in bringing their professional skills to bear on the grave public issues in which science is involved. They must not be seen as leading sequestered and comfortable lives indifferent to the great problems of the world and of their own countries, but as contributing some of their time and energy and expertise to their resolution. They are in a unique

> position to appreciate early the problems, the dangers and the advantages likely to follow from scientific developments and to make their findings known to governments and to peoples.

As has been clarified earlier, until the time of the Second World War, science essentially functioned at the fringes of society – the interactions were largely within the scientific community rather than with society. There were exceptions such as the role of scientists in the development of the chemical industry more than hundred years ago; as also in a variety of areas in which science has proved to be hugely beneficial for society such as the applications of X-Rays in medicine, the use of radio waves and broadcasting for communications, the development of chemotherapeutics and antibiotics which had a major impact on bringing down death rates and other such examples. It was, however, during the Second World War and, thereafter, that many aspects of science have come about that have large public implications, and which figure extensively in the audio-visual and print media.

The first of these was undoubtedly the development of the atomic bomb and the concept of weapons of mass destruction. This led to the famous Bertrand Russell – Albert Einstein manifesto and the creation of the Pugwash Movement. This was essentially a warning by scientists to society about the grave threats that such weapons posed. More specifically, scientists played a role in bringing to the notice of the governments the production of strontium 90 in nuclear explosions – and when these took place as atmospheric tests, that this isotope would be washed down and, being similar to calcium, find its way, through the food chain of plants and animals, into human bodies where it would produce radioactive damage. Whilst the atmospheric tests were important to the nation conducting them, this unintended side-effect would be global. This did have an impact; it led to a Treaty on Prohibition of Atmospheric Testing of Nuclear Weapons.

Since then, the number of grave issues faced by society has increased manifold. It would not be meaningful to list all of these. However, a few are indicated, where science has played an important role in bringing the issues to the notice of governments, and to the people of the world – demonstrating how prescient was Cecil Powell's remark in 1968.

There was the discovery of the ozone hole in the atmosphere by the scientists working in the Antarctica. This was essentially due to the use of ozone-damaging chloro-fluoro carbons and related chemicals by industry.

The destruction of the ozone layer (or a significant damage to it) would result in harmful ultra-violet radiation reaching the surface of the earth, where it can induce cancer. This was an event of grave societal consequence brought to the notice of governments by science. It also involves a straightforward technological fix: how to develop and introduce CFC substitutes into a global economy, which would do no damage to the ozone layer, and be economically acceptable. Through a series of intergovernmental meetings and treaties, efforts on this front have moved forward satisfactorily. Further, the ozone concentrations are also regularly monitored to keep track of new potential hazards.

Another area of long-term and grave concern to society is the increase in the concentration of carbon-dioxide and other greenhouse gases (particularly methane) in the atmosphere. At the time the Industrial Revolution had taken off, around 1780, with the development of the steam engine, CO_2 levels in the atmosphere were around 280 parts per million (ppm). They have reached over 380 ppm today with a slow rise at first and since then accelerating. If we adopt a Business As Usual (BAU) scenario, the levels could go up to 550 or even 700 ppm. Whilst we can measure the greenhouse gas concentrations with a degree of certainty, one cannot say the same about their future concentrations, pathways, sinks, and budgets at various stages. This involves an understanding of many aspects of the earth system, and their feedback interactions, particularly the inertia of the oceans, in acting as 'sinks' as also responding to temperature rise. An increase in greenhouse gas concentration would cause global warming. The Inter-governmental Panel on Climate Change, which won the Nobel Prize in 2007 for its sustained work over three decades in this area, has concluded that the warming would be in the range of 1.4-5.8 degrees Celsius. Whilst this may be small in relation to the range of temperatures encountered over the globe, it represents a global average warming with many impacts. It will be noticed that the warming has been indicated as being over a broad range and lacks the certainty associated with science. Further, as one proceeds to various impacts, such as melting of polar ice caps and of glaciers, change of surface reflectivity, extreme events in precipitation, sea level rise and the like, the uncertainties keep increasing. In such cases, the attitude in society and of many governments is that: if science is so uncertain about the magnitude of these impacts, why not wait till we know better. This is particularly so, since any measures to prevent such changes will involve behavioural change in society, with changes in lifestyle and increased costs. One is here dealing with non-linear processes that are potentially catastrophic. The potential impacts, therefore, need to be included in risk assessments. This is where there can be a prediction of the direction and possible magnitude of climate change but with considerable uncertainty concerning its details.

The subject of climate change is essentially the result of work of the Inter-governmental Panel on Climate Change and scientific work initiated, undertaken and managed particularly through the World Climate Research Program organized jointly by the World Meteorological Organization (WMO) and the International Council for Science (ICSU). It was a subject of major debate in the Second World Climate Conference held in Geneva in 1990, and became the basis for recommendations to the UN Conference on Environment and Development (UNCED), held in Rio in 1992. This resulted in the UN Framework on Climate Change, leading to the Kyoto Protocol and its targets in 1997. In spite of what science has done to forewarn Governments, the most recent report of IPCC is most worrisome.

The world scientific community has, therefore, fulfilled, in some sense, the role that Cecil Powell had indicated for it, in bringing to the notice of governments, and to peoples, some of these grave issues.

Another important subject that was brought to public notice by the scientific community at the Rio Conference in 1992, and prior to that, related to the loss of biological diversity. This has been a subject discussed in a variety of scientific meetings and, particularly, by large international non-governmental scientific organizations. The International Council for Science (ICSU) had, in the mid 1980s, initiated a program known as IGBP (International Geosphere Biosphere Program). This was essentially because it was felt by scientists that one could not discuss aspects relating to the geosphere (atmosphere, the cryosphere, oceans, climate change, and the like) without simultaneously taking note of the biosphere and the interactions that exist between these major domains.

We are aware that there is enormous diversity of other living matter (plant and animal kingdoms) with which we share our planet. Whilst we may have identified and named around one and half million, there are probably, in existence, anything from 3 to 15 million species. It is the insects and bacteria and less complex organisms concerning which we have very little knowledge. This bio-diversity, consisting of plants and animals, is vital for our existence; a few plants of economic value have been identified and form the basis for our food security today. A large part of our pharmaceuticals come from the plant kingdom.

And yet, as human beings multiply, and populations grow, the demand for land increases – for agriculture, plantations, industry, urbanisation, infrastructure and a variety of such purposes. There has been large scale unsustainable exploitation of forest areas, particularly from the viewpoint of the resources that they provide. This has been significantly aggravated by the nature of the consumer society that has been developing, which has been unmindful of the damage done to the eco-system. Scientists have, therefore, been deeply concerned about the loss of bio-diversity and its consequent implications. This was brought to the fore at the Rio Conference in 1992. It resulted in a convention in biodiversity; and also another on forest principles (which was non-binding).

There are many other issues with grave implications for society today, such as: the spread of new and emerging diseases, particularly with crossover infections from animal systems to people, the new opportunities being made available to society through stem cell work, genetic engineering, cloning, and the like, all of which have significant, long-term ethical implications. There are also broad areas of environment and ecology where society seldom looks at the price being paid for certain pathways of development that it has adopted. For example, what is the price to be paid to the ecological services provided by water. It is estimated to run into trillions of dollars and yet given no importance in discussions on development.

In all of these areas, we are dealing with non-linear systems, which can become, very rapidly, far-from-equilibrium systems. Whilst one can make certain predictions about the directions in which these developments can take place, it would be difficult to make unambiguous, clearcut, predictions on where we will get to, and its consequences. Earth system science is in its infancy to be able to make exact predictions and, furthermore, to assert with any degree of certainty.

In the meantime, decisions have to be taken by governments and by society. For this, advice from the scientific community will be called for. This would involve the use of the precautionary principle in many cases to avoid getting into pathways that might lead to catastrophic events. But, it would also necessitate increased public understanding concerning predictability and how certain science can be. This is the role that scientists will have to play.

CHAOS IN SELF-EXCITING DYNAMOS AND THE MAIN GEOMAGNETIC FIELD

RAYMOND HIDE

Summary

Nonlinear feedback and coupling (F&C) in dynamical systems operating under fixed boundary conditions usually produce chaotic fluctuations, which are disorderly and unpredictable beyond a finite 'predictability horizon' ([(Note) A], [(References) 1-5]). But in some circumstances nonlinear F&C *inhibit* chaos, producing *order* rather than *disorder.* Both types of behaviour may have to be invoked in the interpretation of long-term variations of the main geomagnetic field (MGF). This is generated by (mainly buoyancy-driven) magnetohydrodynamic (MHD) flow in the Earth's liquid metallic outer core, where the electrical conductivity is high enough (but not too high) for efficient MHD self-exciting dynamo action to take place. Amongst the nonlinear F&C agencies operating in the MHD 'geodynamo' are Lorentz forces, involving interactions between the electric currents generated by the geodynamo and concomitant magnetic fields. One generic nonlinear process in self-exciting dynamos is the redistribution of kinetic energy by such forces. When operating within a self-exciting Faraday-disk homopolar dynamo with its coil loaded with a nonlinear series motor, Lorentz forces can cause persistent large-amplitude chaotic fluctuations. But over a wide range of conditions they inhibit, rather than promote, persistent fluctuations, in some cases eliminating them altogether. If this 'nonlinear-quenching' process occurs in the MHD geodynamo it could account for the high degree of intermittency seen in the long-term behaviour of the MGF, as exhibited by the time-series of geomagnetic 'polarity reversals', the most striking features of which include intervals lasting as long as 3x10^7 years during which no polarity reversals appear in the palaeomagnetic record. Implied by this hypothesis, which could be tested with the aid of

valid numerical geodynamo models, is that eddies driven mainly by Lorentz forces play a crucial role in the attenuation of fluctuations. Also crucial, and in principle testable, are the roles played in the reversal process by modest changes in the lateral boundary conditions imposed on core motions, including those associated with very slow irregular convection in the highly-viscous overlying mantle, and also by the slow increase in size of the underlying solid inner core.

1. Introduction

When the MGF undergoes an occasional reversal in polarity – or, more frequently, a large-amplitude 'excursion' in the orientation of its magnetic axis which fails to achieve a full polarity reversal – it does so in a few thousand years. Determinations by palaeomagnetic workers of the 'fossilized' magnetization of sedimentary and igneous rocks also indicate that many hundreds of polarity reversals, and a larger number of excursions, may have occurred since the Earth came into existence some 4600 million years (Ma) ago – the most recent being the 'Brunhes-Matyama' reversal of *ca*. 0.8 Ma ago, when the MGF dropped in strength to less than about a fifth of its pre-transition and post–transition value [6, 7].

Intervals between reversals range in duration from *ca*. 0.25 Ma ('subchrons') to *ca*. 30 Ma ('superchrons'), the average duration being *ca*. 1 Ma. During the past 400 Ma, covering the geological periods studied most intensively by palaeomagnetic workers, there have been two superchron intervals – the Permian superchron from *ca*. 290 to *ca*. 260 Ma ago, when a magnetic compass would have pointed roughly south, and the Cretaceous superchron from *ca*. 110 to *ca*. 80 Ma ago, when the polarity was the same as it is today – and there may have been many such intervals at earlier times.

The detailed interpretation of polarity superchrons and other features of the long-term behaviour of the MGF in terms of basic processes occurring at great depths within the Earth ranks as a major problem in geophysics. Indeed, the reversal time-series may hold valuable and singular clues to the structure, dynamics and evolution of the Earth's metallic core and overlying mantle. The MGF is a manifestation of more or less irregular motions at speeds of a fraction of a millimetre per second (kilometres per year) in the liquid metallic outer core, as evinced by detailed observations of the MGF over the past few centuries [B]. Few now disagree that the MGF must be due to electric currents generated and maintained in the outer core by a MHD self-exciting dynamo process involving inductive interactions between the magnetic field and core motions – motional induction (rather than chemical or thermoelectric effects) being the only quantitatively-viable agency for providing the electromotive forces needed to maintain the electric currents against ohmic resistance [7-12].

The near alignment of the Earth's magnetic axis with its rotation axis – a property exploited in the navigational use of the magnetic compass [B] – is most probably due to the dominant influence of gyroscopic (Coriolis) forces on core motions [13]. Less obviously, but importantly, such forces would also render the spatial and temporal characteristics of the MGF sensitive not only to the presence of the solid electrically-conducting inner core but also to modest lateral variations in the (thermal, mechanical and electromagnetic) boundary conditions imposed on motions in the fluid outer core by the overlying highly-viscous mantle [14, 9]. As a working hypothesis the latter suggestion has been taken seriously by geophysicists for the past few decades – from the time when dynamical models of the Earth's mantle in which convective flow occurs at speeds with which continents drift apart (centimetres per year) and extends throughout the whole depth of the mantle, down to the core-mantle boundary (CMB), became generally acceptable [15, 16].

Mantle convection varies on geological timescales and concomitant variations in the boundary conditions imposed by the mantle on motions in the underlying outer core would influence the degree of intermittency exhibited by the observed timeseries of polarity reversals. Some degree of intermittency would also be associated with any intrinsic chaotic behaviour of an essentially nonlinear geodynamo operating under *fixed* boundary conditions. The unexpected discovery of 'nonlinear quenching' of chaos in a physically-realistic Faraday-disk self-exciting dynamo loaded with a nonlinear series motor ([17], see also Section 3 below) indicated a useful line of research towards the physical interpretation of the intermittency seen in the reversals time-series, in terms of the influence of time-varying boundary conditions on chaos in the geodynamo [9].

Some of this geophysical contribution to a Pontifical Academy of Sciences symposium on 'Predictability in Science: Accuracy and Limitations' is based on background material prepared originally for informal seminars on the dynamics and MHD of spinning fluids. The References and Notes given below indicate sources of diagrams and tables used as visual aids (but which for reasons of space cannot be included here) and of useful technical bibliographies.

2. Self-Exciting Dynamos and Cosmical Magnetic Fields

The first to outline the idea of the self-exciting dynamo process in a fluid was J. Larmor when, in 1919, he argued that the magnetic fields of sunspots are produced and maintained by motional induction, involving thermal convection in the outer layers of the Sun. The dynamo process is now widely invoked in the interpretation of cosmical magnetic fields, i.e. those of galaxies, stars and planets, including the MGF supposedly produced by the 'MHD geodynamo' operating in the Earth's liquid metallic outer core [9,10,12,15,16,19-22].

Such dynamos are governed by the four-dimensional (space and time) nonlinear partial differential equations (PDEs) of MHD. These express the laws of electromagnetism, mechanics and thermodynamics, to be solved under realistic boundary conditions [10-12]. Dynamo theory aims to understand the complex details of the processes involved. During the past half-century, since the publication by G.E. Backus and A. Herzenberg of the first mathematical 'existence proofs' based on the (pre-Maxwellian) equations of electromagnetism [8-12; C], dynamo theory has developed rapidly as an active area of applied mathematics. In common with other areas of geophysical and astrophysical fluid dynamics, dynamo theory now benefits from the availability of powerful computers, capable of tackling, with growing success, the full set of equations [10,12]. But the subject suffers not only from paucity of observations but also, in comparison with other areas of fluid dynamics, from lack of guidance from crucial laboratory experiments. Interesting laboratory work has certainly been undertaken [10], but with the electrically-conducting fluids available for use on the small scale of the terrestrial laboratory it is not possible to carry out wide-ranging investigations of the MHD processes thought to be involved in self-exciting dynamos.

The most extensive mathematical studies to date are those of 'kinematic' dynamo models, in investigations of which, for reasons of mathematical expediency, the governing equations are simplified by specifying *ab initio* the fluid velocity field in the equations of electromagnetism [C; 8,10,11], thereby removing nonlinearities from the equations and obviating the need to solve them simultaneously with the equations of mechanics and thermodynamics. Such studies can be important, especially in the elucidation of spatial structure. But they shed little light on temporal behaviour, in the investigation of which it is necessary to simplify the equations in other ways, as in some of the successful 'mean-field' dynamo models, where some spatial details are parameterized rather than represented explicitly [9,10,12].

Further simplifications are needed when interest centres on the essentially nonlinear processes that underlie the complex large-amplitude fluctuations seen in observational data, such as the time-series of polarity reversals and excursions of the MGF. Recourse then has to be made to analyses of what mathematicians term 'low-dimensional' models (and physicists term 'toy' models). These are governed by nonlinear ordinary differential equations (ODEs)(rather than PDEs), requiring only modest computing facilities for their investigation. But great care has to be taken when formulating a model to ensure that it is not oversimplified to the point of being physically unrealistic and mathematically misleading.

The most extensively-studied low-dimensional self-exciting dynamo models are based on the Faraday-disk homopolar dynamo. (For references to an extensive literature starting nearly half a century ago with pioneering studies by E.C. Bullard, T. Rikitake, W.V.R. Malkus, H.K. Moffatt and others, see [10, 18, 23]). Some models are physically unrealistic, usually because mechanical friction is neglected thereby rendering the governing ODEs 'structurally unstable'. But detailed analyses of the 'structurally-stable' ODEs governing physically-realistic models are of interest not only in their own right but also because they offer insights capable of guiding research into the much more complex MHD systems [17].

3. Physically-Realistic Self-Exciting Dynamo Models

Self-exciting dynamos are electromechanical engineering devices or naturally-occurring MHD fluid systems that are capable of converting mechanical energy into magnetic energy without the aid of permanent magnets. They differ widely in their details but they all share the following essential characteristics [18; 23; C]:

(a) motional induction – as represented in the PDEs governing MHD dynamos by the nonlinear term $\boldsymbol{u}\mathrm{x}\boldsymbol{B}$ (where $\boldsymbol{u}$ is the (Eulerian) flow velocity at a general point P and $\boldsymbol{B}$ is the magnetic field, see equations (C1 & C5)) – is responsible for converting mechanical energy into magnetic energy, which starts with the amplification of any infinitesimally-weak adventitious background magnetic field;

(b) motional induction must overcome ohmic losses for amplification to occur, implying that the electrical resistance of the system must be sufficiently low – i.e. a sufficiently *high* magnetic Reynolds number $R{=}UL\mu\sigma$ in MHD dynamos (see equation (C7), also [7-10]), where U is a characteristic flow speed, L a characteristic length, σ the electrical conductivity of the fluid and μ its magnetic permeability;

(c) but the electrical resistance must not be so low that the magnetic field is unable to diffuse beyond the dynamo region, which sets an *upper* limit on R in MHD dynamos;

(d) Lorentz forces – as represented in governing equations of MHD dynamos by the nonlinear term $\boldsymbol{j}\mathrm{x}\boldsymbol{B}$, where $\boldsymbol{j}$ is the electric current density at P – redistribute kinetic energy within the system, thereby retarding buoyancy-driven eddies and accelerating motions in other parts of the eddy spectrum;

(e) mechanical friction – viscosity in MHD dynamos – no matter how weak is never negligible;

(f) internal F&C – as represented by the terms $\boldsymbol{u}\mathrm{x}\boldsymbol{B}$ and $\boldsymbol{j}\mathrm{x}\boldsymbol{B}$ in MHD dynamos – give rise to behaviour characteristic of nonlinear systems, including sensitivity to initial conditions, non-uniqueness, chaotic large amplitude fluctuations, hysteresis, nonlinear stability, etc.

A simple physically-realistic low-dimensional model (that takes mechanical friction into account and includes a crucial circuit element which enables Lorentz forces to redistribute kinetic energy) comprises a Faraday disk-and-coil arrangement with a series electric motor loading the coil [17,18,23]. The disk – to the axle of which the stationary coil is connected by a sliding contact and to the rim of which the series motor loading the coil is connected by another sliding contact – is driven into rotation with (dimensionless) angular speed $y(t)$ by a steady applied couple proportional to the dimensionless parameter a (say, see equation (3.4)), which is inversely proportional to the moment of inertia of the disk. Here the dimensionless time t is measured in units of the ratio of the self-inductance of the coil to the total electrical resistance of the dynamo circuit – the corresponding timescale in the case of an MHD dynamo being $\mu\sigma L^2$, which is several thousand years for the geodynamo. Retarding the motion of the disk are a frictional couple $-ky(t)$ and a Lorentz couple $-ax(t)w(t)$, where $x(t)$ is the main electric current generated by the dynamo and $w(t)$ is the magnetic flux intersecting the disk. In the absence of Lorentz forces, when friction alone retards the motion of the disk, y has the steady value a/k.

The armature of the series motor is driven into rotation with angular speed $z(t)$ (relative to the stationary ambient magnetic field within the motor) by a Lorentz couple $x(t)f(x(t))$ produced by the dynamo current $x(t)$, and it is retarded by a frictional couple $-lz(t)$ (see equation (3.5)). Here

$$f(x(t)) = 1 - e + esx(t), \qquad 0 \leq e \leq 1, \tag{3.1}$$

where the value of the crucial parameter e depends on the design of the motor. The second and first terms on the right hand side of this equation are in the ratio $esx(t)$ to (1-e), so the parameter e is a measure of the nonlinearity of the electromechanical characteristics of the motor, the linear case being when $e = 0$. The nonlinear contribution to f, proportional to $esx(t)$, is produced by diverting the dynamo current through the stationary field windings of the motor (s being a measure of the mutual inductance between the armature and the field windings). This nonlinear 'internal' contribution to f is complemented by a linear contribution, proportional to (1–e), provided by 'outside' sources of the ambient magnetic field in which the armature rotates. In general these sources comprise (i) currents generated by 'secondary' dynamos with e=1, and (ii) permanent magnets, but the latter must be excluded from consideration when dealing with systems of interest as low-dimensional models of natural MHD dynamos [23].

The nonlinear autonomous set of ODEs satisfied by the time-dependent variables (x, y, w, z) is the following [18]:

$$\mathrm{d}x/\mathrm{d}t + m\mathrm{d}w/\mathrm{d}t = (1 + m)\,[-x + yw - bzf(x)], \tag{3.2}$$

$$n\mathrm{d}w/\mathrm{d}t = (x - w), \tag{3.3}$$

$$\mathrm{d}y/\mathrm{d}t = a(1 - xw) - ky, \tag{3.4}$$

$$\mathrm{d}z/\mathrm{d}t = xf(x) - lz, \tag{3.5}$$

Equations (3.2 & 3.3) express Kirchhoff's laws applied respectively to the dynamo current, $x(t)$, flowing in the main circuit and the induced azimuthal 'eddy' current (proportional to $w(t)$) flowing in the disk. Equations (3.4 & 3.5) express angular momentum considerations applied respectively to the motion of the disk and the motion of the armature of the motor.

The essentially non-negative dimensionless control parameters (a, k, n, m) specify the electromechanical characteristics of the disk, while (b, l, s, e) specify those of the motor. The parameter n is inversely proportional to the electrical resistance of the disk to the flow of the azimuthal eddy current; m is proportional to the square of the mutual inductance between the disk and coil and inversely proportional to the difference between the product of the self inductances of the disk and coil and the square of their mutual inductance; and b is proportional to the self inductance of the armature of the motor and inversely proportional to its moment of inertia. The F&C terms in these governing equations are $(1+m)(yw-bzf(x))$ in (3.2), $-axw$ in (3.4), and $xf(x)$ in (3.5).

Instability analyses of the equations combined with numerical integrations using both digital and analogue computers show that when (i) a/k is large enough for dynamo action to occur at all, and (ii) the parameter e is not too close to unity, over wide ranges of conditions – as specified by the other control parameters – their solutions exhibit chaotic behaviour, in which the dynamo current $x(t)$ undergoes large amplitude fluctuations reminiscent of polarity excursions and reversals of the MGF reversals [17, 18, 10]. Remarkably, however, chaotic behaviour is rare when e is close to unity, with fluctuations disappearing altogether when e=1, irrespective of the values of the other control parameters! It is this result, as we shall see in Section 4, that offers a possible basis for interpreting the long-term behaviour of the MGF, with its highly intermittent time-series of polarity reversals.

The original discovery of 'nonlinear-quenching' of persistent fluctuations (after the decay of initial transients) was made during an investigation of cases when n=0 [17]. The discovery emerged from the unexpected finding that although 'Hopf bifurcations' can occur when $0 \leq e < 1$, they disappear altogether when e=1, irrespective of the values of the other control parameters. The phenomenon may turn out to be fairly general, for it was subsequently found in solutions of other autonomous sets of nonlinear ODEs [24], obtained by taking sets (e.g. Lorenz, Rössler) that are known to possess persistent chaotic solutions and modifying their F&C terms [24], along lines suggested by the original work reported in [17].

We note that for every solution (x, y, w, z) of equations (3.1-3.5) there is a corresponding solution, ($-x$, y, $-w$, z), with the same motions, (y, z), but oppositely-directed currents. This 'magnetic symmetry' – of interest when considering 'reversals' of the MGF (see Section 4 below and [C]) – is obvious by inspection when e=1. But its demonstration (not presented here) for other (lower) values of e requires detailed considerations of the currents in the secondary dynamo to which the primary dynamo has to be coupled in order to achieve (without using permanent magnets) values of e that are less than unity, thereby stimulating chaotic large-amplitude fluctuations in $x(t)$.

4. Long-Term Variations in the Main Geomagnetic Field

Very slow intermittent convection taking place within the Earth's highly-viscous mantle that overlies the liquid metallic outer core must give rise to fluctuations, on geological timescales, in the lateral variations of the physical and chemical conditions prevailing at the core-mantle boundary (CMB) [14, 15, 9, 25]. Owing to Coriolis forces, these cause marked changes in both spatial and temporal characteristics of core motions, and consequently in the magnetic fields they produce. So it is not unreasonable to suppose that each polarity superchron could be associated with a long quiescent period, when convection in the lower mantle (but not necessarily in the upper mantle) is comparatively feeble [18].

Because the CMB would then be relatively undisturbed, the corresponding state of the geodynamo (on this particular hypothesis) would be highly stable owing to processes analogous to the nonlinear quenching mechanism found in disk dynamos. What corresponds within the core to the presence of a series motor in the disk dynamo with a Lorentz torque proportional to the square of the electric current (the case when $e=1$, see equation (3.1)) is that part of the spectrum of core motions containing eddies driven mainly by Lorentz forces, rather than by buoyancy forces. However, during comparatively intensive phases of intermittent convection in the lower mantle conditions at the CMB become disturbed – with buoyancy forces associated with increased lateral temperature gradients in the lower mantle now contributing to the driving of core motions and distortions in the shape and other conditions prevailing at the CMB also influencing the pattern of motions. Concomitant distortions and other changes of the magnetic field within the core would stimulate changes in the geodynamo, possibly placing it within regimes corresponding to those found in a single-disk dynamo when the parameter e is no longer close to unity, which favour frequent reversals and excursions.

A simple model capable of such behaviour comprises a ‘primary’ single-disk dynamo subject to *steady* forcing and interacting with a ‘secondary’ single-disk dynamo subject to *slowly varying* irregular forcing. The secondary dynamo interacts with the primary dynamo solely by modulating the ambient magnetic field in which the armature of the series motor of the primary dynamo rotates, thereby modulating the value of e in the primary dynamo and moving the system slowly between the quenched and chaotic regimes, in response to the modulated forcing of the secondary dynamo.

Typical time-series of the dynamo current would include ‘superchrons’ – when x is steady and either positive or negative depending on initial conditions – occurring during intervals when the influence of the secondary dynamo on the primary dynamo is so weak that e is close to unity in the primary dynamo. By contrast, highly time-dependent behaviour found between ‘superchrons’ occurs during phases when the current in the secondary dynamo is strong enough to reduce the value of e in the primary dynamo to values significantly less than unity.

Noteworthy features of typical time-series of $x(t)$ include fluctuations that are more frequent than reversals, less pronounced in amplitude, and exhibit systematic monotonic build up in amplitude before each reversal occurs. They are reminiscent of the excursions seen in the geomagnetic record which, according to the ideas presented here, should be less pronounced during superchrons than at other times.

5. Concluding Remarks

Modern research on the MGF includes improved analyses and applications of geomagnetic data derived from ground based and orbiting satellite observations. In rock magnetism, time-series of polarity reversals and excursions continue to be extended and refined, and in theoretical work, applied mathematicians interested in MHD continue to investigate and apply the equations governing self-exciting dynamos. Numerical modelling of dynamos [10,12] has strengthened collaboration between observational workers and theoreticians, but the MGF is a complex phenomenon and much remains to be elucidated. The main ideas outlined in this paper were put forward in the hope that they will stimulate crucial observational and theoretical studies by those equipped to carry them out.

REFERENCES

1. Lighthill, J., 'The recently recognized failure of predictability in Newtonian mechanics', *Proc. R. Soc. Lond.* A407, 35-50 (1986).
2. Lorenz, E.N., *The essence of chaos,* London: UCL Press (1993).
3. Bjerknes, V., 'Das Problem der Wettervorhersage, betrachtet vom Standpunkt der Mechanik und Physik', *Meteorol. Zeitschr*. 21, 1-7 (1904).
4. Richardson, L.F., *Weather prediction by numerical process,* Cambridge University Press (1922).
5. Lorenz, E.N., 'Deterministic non-periodic flow', *J. Atmos. Sci*. 20, 130-141 (1963).
6. Gubbins, D., Kent, D. & Laj, C. (eds.), 'Geomagnetic polarity reversals and long term secular variation', *Phil. Trans. R. Soc. Lond.* A358, 889-1223 (2000); Opdyke, N.D. & Channell, J.E.T., *Magnetic stratigraphy,* San Diego: Academic Press (1996); Jacobs, J.A., *Reversals of the Earth's magnetic field*, Cambridge University Press (1994).
7. Merrill, R.T., McElhinny, M.W. & McFadden, P.L., *The magnetic field of the Earth*, Academic Press (1996).

8. Backus, G., Parker, R. & Constable, C., *Foundations of geomagnetism,* Cambridge University Press (1996).
9. Gubbins, D. & Herrero-Bervera, E. (eds.), *Encyclopedia of geomagnetism and paleomagnetism*, Springer Geosciences (2007).
10. Dormy, E. & Soward, A.M. (eds.), *Mathematical aspects of natural dynamos*, Grenoble Science Publishers (2007).
11. Friedlander, S. & Serre, D. (eds.), volume 2 of *Handbook of mathematical fluid dynamics*, Elsevier Science BV (2003); Ferriz-Mas, A. & Nuñez, M. (eds.), *Advances in nonlinear dynamos*, London: Taylor & Francis (2003); Moffatt, H.K., *Magnetic field generation in electrically-conducting fluids*, Cambridge University Press (1978).
12. Glatzmaier, G.A. & Roberts, P.H., 'Simulating the geodynamo', *Contemp. Phys*. 38, 269-288 (1997); Zhang, K.K. & Jones, C.A., 'The effect of hyperviscosity on geodynamo models', *Geophys. Res. Lett*., 24, 2869-2972 (1997); LeMouël, J.-L., Allègre, C.J. & Narteau, C., 'Multiple scale dynamo', *Proc. Nat. Acad. Sci. USA* 94, 5510-5514 (1997); Busse, F.H., 'Homogeneous dynamos in planetary cores and in the laboratory', *Annua. Rev. Fluid Mech*. 32, 383-408 (2000); Dormy, E., Valet, J.-P. & Courtillot, V., 'Numerical models of the geodynamo and observational constraints', *Geochemistry, geophysics*, *geosystems*, Paper number 2000GC 000062, ISSN 1525-2027, 42 pages (2000); Narteau, C. & LeMouël, J.-L., 'Transient evolution regimes in a multiscale dynamo model: timescales of the reversal mechanism', *J. Geophys. Res.* 110, B01, doi; 1029/2004J B002, 983, (2005).
13. Elsasser, W.M., 'The origin of the Earth's magnetic field', *Phys. Rev.* 55, 489-498 (1939).
14. Hide, R., 'Motions of the Earth's core and mantle and variations of the main geomagnetic field', *Science* 157, 55-56 (1967); 'Interaction between the Earth's liquid core and solid mantle', *Nature* 222, 1055-1056 (1969); 'On the Earth's core-mantle interface', *Quart. J.R. Meteorol. Soc*. 96, 579-590 (1970); 'On the role of rotation in the generation of magnetic fields by fluid motions', *Phil. Trans. R. Soc. Lond.* A306, 223-234 (1982).
15. Jones, C.A., Soward, A.M. & Zhang, K.K. (eds.), *Earth's core and lower mantle*, London: Gordon & Breach (2003).
16. Melchior, P.J. *The physics of the Earth's core,* Oxford: Pergamon Press (1986).
17. Hide, R., 'Nonlinear quenching of current fluctuations in a self-exciting homopolar dynamo', *Nonlinear Processes in Geophysics* 4, 201-205

(1997); 'Structural instability of the Rikitake disk dynamo', *Geophys. Res. Letters* 22, 1057-1059 (1995).

18. Hide, R., 'Generic nonlinear processes in self-exciting dynamos and the long-term behaviour of the main geomagnetic field, including polarity superchrons', *Phil. Trans. R. Soc. Lond.* A358, 943-955 (2000).
19. Proctor, M.R.E. & Gilbert, A.D. (eds.), *Lectures on solar and planetary dynamos*, Cambridge University Press (1994).
20. Soward, A.M., Jones, C.A., Hughes, D.W. & Weiss, N.O. (eds.), *Fluid dynamics and dynamos in astrophysics and geophysics*, London: Taylor & Francis (2004).
21. Thompson, M.J. & Christensen-Dalsgaard, J. (eds.), *Stellar astrophysical fluid dynamics*, Cambridge University Press (2003).
22. Parker, E.N., *Cosmical magnetic fields: their origin and their activity*, Oxford: Clarendon Press (1979); Krause, F. (ed.), *The cosmic dynamo*, Dordrecht: Kluwer (1993); Mestel, L., *Stellar magnetism*, Oxford: Clarendon Press (2003); Schrijver, C.J. & Title, A.M., *Solar and stellar magnetic activity*, Cambridge University Press (2000).
23. Hide, R., 'The nonlinear differential equations governing a hierarchy of self-exciting coupled Faraday-disk homopolar dynamos', *Phys. Earth Planet. Interiors* 103, 281-291 (1997).
24. Hide, R., McSharry, P.E., Finlay, C.C. & Peskett, G.D., 'Quenching Lorenzian chaos', *Int. J. Bifurcation and Chaos* 14, 2875-2884 (2004).
25. Gunis, M., Wysession, M.E., Knittel, E. & Buffett, B.A., (eds.), *The core-mantle region,* Washington DC: American Geophysical Union (1998).

NOTES

A. Predictability Horizons

> ...notwithstanding the continuing success of applications of Newtonian dynamics in many fields of science and engineering...modern theories of dynamical systems have...clearly demonstrated that *the equations of Newtonian dynamics do not necessarily exhibit the 'predictability' property*. Indeed, ...recent researches have shown that in wide classes of...simple systems satisfying the equations *predictability is impossible beyond a definite time horizon* (J. Lighthill, President of the International Union of Theoretical and Applied Mechanics [1]).

The idea of a 'predictability horizon' is important [1,2]. Nearly fifty years ago meteorologists, encouraged by advances in computing and satellite technology, started planning their ambitious 'Global Atmospheric Research Programme' (GARP), which was formally announced in 1963 (in a speech by the then President of the United States, J.F. Kennedy). The planners hoped that the predictability horizon for useful numerical weather forecasts based on greatly improved data sets and advanced computational methods could eventually be extended from a day or so to more than a week. Computer technology would render feasible methods based on the governing PDEs of hydrodynamics along lines first discussed decades earlier by V. Bjerknes, a Norwegian meteorologist (and later a member of the Pontifical Academy of Sciences), and L.F. Richardson, a British meteorologist [3,4]. From their experience with weather forecasting and, more recently, climate forecasting, modern meteorologists appreciate – at least as well as other geophysical scientists engaged in research on large-scale natural phenomena – the difficulties encountered when translating understanding of basic dynamical processes into practical schemes for predicting future behaviour. Indeed, it was no accident that one of the major advances in what in 1973 became known as 'chaos theory' was the publication ten years earlier, in a leading meteorological journal, of a paper [5] entitled 'Deterministic non-periodic flow' by E.N. Lorenz, who had been engaged with other leading meteorologists in the assessment of likely improvements in 'predictability horizons' to be expected from GARP. The Earth's atmosphere is by far the most intensively-studied 'geophysical fluid', and many ideas about predictability stem from research connected with the very difficult practical activity of forecasting the atmosphere's future behaviour.

B. Geomagnetic Secular Variation and Core Motions

Geomagnetism is a major branch of modern geophysics. It emerged as a science after centuries of navigational use by seafarers of the magnetic compass, the invention of which has been compared in its historical importance to that of gunpowder and printing. The magnetic compass is not a perfect instrument, for its magnetized needle deviates in alignment from true North (based on the Earth's rotation axis) at a typically small but significant 'declination' angle, usually denoted by *D*. According to the 1980 global chart showing isolines ('isogonals') of *D,* the 'agonic' line (where *D*=0) then passed through North America and South America and also through Australia, Malaysia and the Western Pacific; D was generally negative in the longitudes of Europe, Africa and Asia, positive over most of the Pacific Ocean, and typically less than 20° in magnitude over most of the Tropics and mid-latitude regions.

A key development came with the sixteenth-century discovery in Europe that a magnetized needle allowed to swing freely in a *vertical* as well as a horizontal plane would point below the horizon at the so-called 'dip' or 'inclination' angle I. In 1980 I was roughly +55° in Rome, +65° at the higher latitudes of London and Paris, and negative over most of the geographic southern hemisphere. By definition I is +90° at the 'magnetic north pole' (then, as now, located in northern Canada), –90° at the 'magnetic south pole' (located in Antarctica), and is everywhere zero on the 'magnetic equator'. This closed line in 1980 was confined to the geographic latitude belt extending from *ca*. +15° to *ca*. –15°and cut the geographic equator at two points, one in the Atlantic Ocean and the other in the Pacific Ocean.

Inspired by the sixteenth-century discovery of the magnetic dip, physician W. Gilbert of Colchester undertook systematic experiments on the properties of lodestone and other magnetic materials. His wide-ranging findings were reported in 1600 in *De Magnete* (full title *De Magnete, Magneticisque Corporibus, et de Magno Magnete Tellure* (*Concerning Magnetism, Magnetic Bodies, and the Great Magnet Earth*)), a treatise said to have influenced Galileo and Kepler and which one leading historian of science declared to be the first truly scientific text book (see Gilbert, W., *De Magnete*, Gilbert Club revised English translation, London: Chiswick Press (1900)). Gilbert's conclusion that the agency responsible for aligning the compass needle must reside not in the heavens but within the Earth itself quickly gained acceptance. But the nature of the agency and its exact location remained mysterious until the twentieth century, when geophysicists inferred from seismological and other evidence that the Earth has a metallic (mainly iron) core with a solid inner part and liquid outer part, and eventually became persuaded, mainly on general quantitative grounds, that the MGF must be a manifestation of ordinary electric currents within the core maintained by electromotive forces due to motional induction involving flow in the outer core [7-12].

Research on this 'self-exciting MHD geodynamo' benefited from twentieth-century discoveries in terrestrial palaeomagnetism and planetary science. In palaeomagnetism systematic investigations of fossilized magnetization of sedimentary and igneous rocks revealed (i) that the MGF must have existed since soon after the Earth was formed some 4600 MY ago, and

(ii) (as we have seen in section 1 above) that at irregular intervals ranging in duration from *ca* 0.25MY to *ca* 30 MY there have been many polarity changes, each taking no more than a few thousand years to be accomplished [6,7]. In planetary science ground-based radio-astronomical observations and magnetic data collected near the major planets by space-craft equipped with magnetometers revealed that Jupiter and Saturn possess magnetic fields aligned nearly (anti-) parallel with their respective rotation axes, and that Uranus and Neptune possess magnetic fields aligned at substantial angles – more than 40° – to their rotation axes [10,22].

Hints that the MGF is not a steady phenomenon were already appearing by the time of Gilbert's death in 1603. Accurate measurements showing that *D* in London decreased more or less steadily from its value of +11.25° in 1571 to +4.05° in 1634 provided the first evidence of what later became known as the 'geomagnetic secular variation' (GSV), a spatially and temporally complex global phenomenon characterized by timescales of decades and centuries. These modest but unexpected temporal changes were announced in 1635 by the professor of astronomy at Gresham College, H. Gellibrand, in a treatise entitled *A Discourse Mathematical on the Variation of the Magneticall Needle*. For practical navigational purposes they implied that charts of *D* would have to be revised from time to time and carry predictions indicating future changes. A recent Ordnance Survey map of London indicates that in 2003 the local value of *D* was –3.5° and its time rate of change was +0.15° per year.

Amongst those who became interested in the GSV were E. Halley and K.F. Gauss, both famous in astronomy for impressive predictions based on Newton's theory of orbital motions in the Sun's gravitational field – Halley of the return in 1758 of the comet that bears his name, and Gauss of the orbit of Ceres, the first asteroid to be discovered (in 1801), by applying his now widely-used 'method of least squares' to observations covering a geocentric arc of no more than three degrees. Their efforts in geomagnetism greatly furthered the systematic acquisition and analysis of data, in Gauss's case including measurements of the intensity at a network of stations, thereby laying foundations for modern research on the MGF.

The practice of presenting geomagnetic data in the form of 'contour maps' was invented for the purpose by Halley and goes back to the publication in 1701 of his chart showing isolines of constant *D* for the Atlantic Ocean. The practice of representing the MGF mathematically, as the sum of contributions made by hypothetical multipoles (dipole, quadrupole, octupole, etc.) located at the centre of the Earth, goes back to the publication in 1838 of Gauss's book *Allgemeine Theorie des Geomagnetismus*, in which the magnetic field is expressed as the gradient of a potential function expanded as an infinite series of spherical harmonics [C].

According to the data thus analyzed [9], in the spherical harmonic expansion of the MGF the dominant term corresponds to the hypothetical centred dipole aligned with the Earth's rotation axis, with an amplitude exceeding those of the equatorial components of the dipole and of all other (i.e. 'non-dipolar') terms in the expansion by at least a factor of five. But it is the 'non-dipolar' part of the MGF that undergoes the most rapid secular changes. On typical magnetic maps, lines of equal annual change of *D* or any other element ('isopors') form sets of oval curves surrounding points at which the changes are most rapid (isoporic foci). A typical set of isopors covers an area of continental size and is separated from neighbouring sets by regions over which changes are comparatively small.

A substantial fraction (about a half) of the GSV can be accounted for in terms of a general westward drift of the non-dipole component of MGF at about 0.18 degrees of longitude per year. The westward drift of the MGF (first noted by Halley from the movement in longitude of the points where the magnetic equator intersects the geographic equator) is a rough measure of typical speeds of fluid motions in the Earth's liquid outer core, somewhat less a millimetre per second (several kilometres per year). There is a general (but not of course detailed) resemblance between global magnetic maps and global maps contouring pressure or other meteorological elements in the Earth's atmosphere, where typical wind speeds are several metres per second. So a century of detailed geomagnetic observations is roughly equivalent, from a dynamical point of view, to no more than a few days of meteorological observations, a useful but hardly excessive data set for detailed scientific studies.

Owing to the twentieth-century decline in the use of the magnetic compass associated with the introduction of new navigational aids such as the gyrocompass and, more recently, the Global Positioning System (GPS) based on signals from a swarm of orbiting artificial satellites, the practical need for determinations and predictions of changes of the MGF has virtually disappeared. But the geomagnetic observations obtained originally for navigational purposes are now exploited by geophysicists undertaking research on the dynamics and structure of the Earth's deep interior. This continuing development has for the past twenty years been facilitated by the International Union of Geodesy and Geophysics under a successful special programme with the acronym SEDI (Study of the Earth's Deep Interior) [6].

The irregular fluctuations of modest amplitude seen in the MGF on GSV timescales, decades and centuries, largely reflect the continual rearrangement (rather than creation or destruction) of magnetic lines of force by the mainly horizontal non-steady motions near the top of the core, in accordance with Alfvén's 'frozen flux' theorem [8; 16; C] (see also Jackson, A., Constable, C.G., Walker, M.R. & Parker, R.L., 'Models of the Earth's magnetic field incorporating flux and radial vorticity constraints', *Geophys. Journ. Inter.*, (submitted); Roberts, P.H. & Glatzmaier, G.A., 'A test of the frozen–flux approximation using a new dynamo model', *Phil. Trans. R. Soc. Lond.* A358, 1109-1121 (2000)). These motions should include a general class of geophysically and astrophysically important 'magnetostrophic' oscillations associated with restoring forces equivalent to the combined effects of Lorentz forces and Coriolis forces when these forces oppose each other. (See Braginsky, S.I., 'Magnetic waves in the Earth's core', *Geomag. & Aeron.* 7, 851-859 (1967); Hide, R., 'Free hydromagnetic oscillations of the Earth's core and the theory of the geomagnetic secular variation', *Phil. Trans. R. Soc. Lond.* A259, 614-647 (1966); also [10, 11, 12, 16]). Any creation or destruction of magnetic field lines occurs on the longer timescales characteristic of processes involving ohmic diffusion and dissipation. Consistent with this picture are polarity reversals and excursions of the MGF, each taking thousands of years [6, 7, 8].

C. Inferences from Equations of Electromagnetism

We assume here a simple spherically-symmetric model of the Earth's interior comprising an electrically conducting 'core' of radius 3485 km surrounded by an electrically-insulating 'mantle' extending out to the surface of the model Earth, radius 6371 km. Underlying the liquid outer part of the conducting core is the solid 'inner core' of radius 1220 km. Observations of the MGF are confined to regions at and near the Earth's surface, but with the aid of the equations of electromagnetism it is possible to make useful inferences about the magnetic field within the Earth, especially at the core-mantle boundary (CMB).

Supposing that the MGF is produced entirely by electric currents within the core we take the magnetic permeability μ to be everywhere equal to that of free space and the electrical conductivity σ to be non-zero and constant throughout the core and zero throughout the mantle. The concepts and ideas developed by nineteenth-century physicists from their discoveries in electromagnetism can be applied to an electrically-conducting fluid such as the Earth's liquid outer core in motion at time t with Eulerian flow velocity $\boldsymbol{u}(\boldsymbol{r}, t)$ at general point P at position $\boldsymbol{r}$ in the chosen reference frame. They give the following nonlinear partial differential equation relating $\boldsymbol{u}$ and $\boldsymbol{B}(\boldsymbol{r}, t)$, the magnetic field, at P:

$$\partial\boldsymbol{B}/\partial t - \mathrm{curl}(\boldsymbol{u}\mathrm{x}\boldsymbol{B}) = (\mu\sigma)^{-1}\,(\mathrm{divgrad})\boldsymbol{B}. \qquad \text{(C1)}$$

This is obtained by eliminating $\boldsymbol{E}$ and $\boldsymbol{j}$ from the equations expressing the (pre-Maxwellian) laws of Gauss, Ampère, Faraday and Ohm, respectively:

$$\mathrm{div}\boldsymbol{B} = 0, \qquad \mathrm{curl}\boldsymbol{B} = \mu\boldsymbol{j}, \qquad \text{(C2, C3)}$$

$$\partial\boldsymbol{B}/\partial t + \mathrm{curl}\boldsymbol{E} = 0, \qquad \boldsymbol{j} = \sigma(\boldsymbol{E} + \boldsymbol{u}\mathrm{x}\boldsymbol{B}), \qquad \text{(C4, C5)}$$

where $\boldsymbol{E}$ is the electric field at P and $\boldsymbol{j}$ the electric current density.

A general result of interest in connection with the interpretation of polarity reversals is that for every solution $(\boldsymbol{u}, \boldsymbol{B})$ of these equations there is a corresponding solution $(\boldsymbol{u}, -\boldsymbol{B})$ if the boundary conditions are independent of the sign of $\boldsymbol{B}$. This 'magnetic symmetry' is also a property of the full MHD equations, for the Lorentz force (per unit volume), $\boldsymbol{j}\mathrm{x}\boldsymbol{B}$, in the equations of mechanics does not change when $\boldsymbol{B}$ changes sign.

According to equations (C2) & (C3), $\boldsymbol{B}$ may be split into two parts, a 'toroidal' part for which the radial component, B_r, is everywhere equal to zero, and which cannot exist in an insulator such as the mantle, and a 'poloidal' part for which B_r is generally non-zero and can exist anywhere. The toroidal part of $\boldsymbol{B}$ (which is associated with the *poloida*l part of $\boldsymbol{j}$) within the conducting core must vanish at the CMB, but its average strength throughout the core is likely to be higher than that of the poloidal part (associated with the *toroidal* part of $\boldsymbol{j}$), implying that much and probably most of the Earth's magnetic energy ($(2\mu)^{-1} B^2$ per unit volume) is associated with field lines that are confined to the core and cannot therefore be observed directly at the Earth's surface [9].

The poloidal part of $\boldsymbol{B}$ possesses field lines that cross the CMB, with a structure which can, in principle, be inferred throughout the insulating 'mantle' from determinations of the MGF at the Earth's surface. The extrapolation procedure makes use of (i) the general *solenoidal* character of $\boldsymbol{B}$ (i.e. $\mathrm{div}\boldsymbol{B} = 0$ everywhere owing to the absence of magnetic monopoles, see equation (C2)) and (ii) its *irrotational* character (i.e. $\mathrm{curl}\boldsymbol{B} = 0$) in insulating regions, where $\boldsymbol{j} = 0$ (see equations (C3) & (C5)). Then it is possible to express $\boldsymbol{B}$ as the gradient of a scalar potential, $-V$ (say), satisfying Laplace's equation $\mathrm{divgrad}V = 0$.

From the solenoidal character of $\boldsymbol{B}$ alone it is possible to infer general topological features of the pattern formed by intersections of the field lines of $\boldsymbol{B}$ with a general closed surface such as a spherical surface S concentric with the centre of our model Earth. Thus, on any S there must always be one or more closed C-lines (null-flux lines) – defined as the loci of points where the normal component of $\boldsymbol{B}$, namely B_r, vanishes, which separate regions where $B_r > 0$ on S from regions where $B_r < 0$. Moreover, the pattern must include one or more pairs of D-points where only the normal component of $\boldsymbol{B}$ is non-zero.In some but not all cases, depending on the behaviour of the non-radial components of $\boldsymbol{B}$ on C-lines and in the vicinity of D-points (features which are related through the 'hairy ball' theorem of topology), there may also be 'touch points' T (say) on C-lines, where the non-radial components of $\boldsymbol{B}$ are tangential to the C-line. And there may be neutral points, N, where all components of $\boldsymbol{B}$ are zero.

Consistent with what would be expected of a field pattern at levels lying well above the source of the field, in 1990 the MGF at the surface of the Earth typically exhibited just one C-line, the magnetic equator, one pair of D-points, the north and south magnetic poles, and no T(touch)-points. The extrapolated field increased in complexity (and strength) with depth and according to one study the corresponding field pattern at the CMB exhibited 4 C-lines 'nested' with the main magnetic equator and 3 'non-nested' C-lines, 13 pairs of D-points additional to the main pair of magnetic poles, and 6 pairs of T- points (Hide, R., Barraclough, D.R. & McMillan, S., 'Topological characteristics of magnetic and other solenoidal vector fields', *Annales Geophysicae* 15 (Supp. 1), page C122, (1997)). The extent to which some of these features are characteristic of magnetic fields produced by self-exciting dynamo action is probably worthy of further study, for we know [8,9,10] that such action cannot maintain a field possessing an axis of symmetry, which would have neither touch points nor non-nested C-lines. When in due course geodynamo computer modellers undertake detailed studies of the long-term behaviour of the MGF, as a useful preliminary they will undoubtedly investigate whether topological features change in characteristic ways during polarity reversals and excursions.

Denote by $N(\mathrm{S}; t)$ the number of intersections of field lines of $\boldsymbol{B}(\boldsymbol{r}, t)$ with S at time t, as given by the surface integral of $|B_r|$ over S (the corresponding surface integral of B_r being zero in virtue of equation (C2)). Determinations of N at the bottom of the mantle based on extrapolations of the surface

MGF using the best geomagnetic data available, including those obtained in 1980 and 2000 from orbiting magnetometers on the Magsat and Oersted artificial satellites respectively, provide convincing evidence that the GSV largely involves the continual re-distribution of field lines (rather than their creation and destruction) by motions in the upper reaches of the core [9; B]. This result is understandable for, according to equation (C1), on timescales much less than $\mu\sigma L^2$ (where L is a characteristic length scale), which is several thousand years when L is comparable with the core radius, the core behaves like a perfect conductor, for which it is impossible to change the total linkage of magnetic field lines, as measured by N. Indeed, a novel method for determining the radius of the electrically-conducting core from magnetic observations alone – by finding that level where N is independent of t – gives values close (to within about 2%) of the accepted value based on the more accurate methods of seismology (see [8,9], also Hide, R., 'How to locate the electrically-conducting fluid core of a planet from external magnetic observations', *Nature* 271, 640-641 (1978)).

The magnetic field in a perfect conductor satisfies Alfvén's 'frozen magnetic flux' theorem

$$\partial \boldsymbol{B}/\partial t - \mathrm{curl}(\boldsymbol{u}\mathrm{x}\boldsymbol{B}) = 0, \qquad \text{(C6)}$$

to which equation (1) tends when the 'magnetic Reynolds number',

$$R = UL\mu\sigma, \qquad \text{(C7)}$$

is very large, U being a characteristic flow speed. On the timescales over which this equation holds within the core the structure of $\boldsymbol{B}$ at the CMB should be such that each C-line, as it moves and suffers distortion under the influence of the horizontal flow just below the CMB, retains both its separate identity and its relationship with dip poles and touch points. Any touch points present move with that flow, but the determination of the flow at other points requires additional information [8,9]. This comes from the equations of motion expressing the laws of mechanics, which show that just below the CMB (where the toroidal part of $\boldsymbol{B}$ is much weaker than at deeper levels within the core), large-scale flow is mainly 'geostrophic', with Coriolis forces in balance with horizontal pressure gradients (as are large-scale flows in the atmosphere and oceans) (LeMouël, J.-L., Gire, C. & Madden, T., 'Motions at core surface in the geostrophic approximation', *Phys. Earth Planet. Interiors* 39, 270-287 (1985)).

Equations (C1) & (C6) lead to one useful general prediction, which we note here in passing. When R is large, Alfvén's theorem (equation (C6)) holds *nearly* everywhere within the fluid. But acceptable mathematical solutions must be such that there exist elsewhere within the fluid localized regions where $\boldsymbol{B}$ changes on length scales that are so much less than L that the right-hand side of equation (C1) is *not* negligible. Otherwise it would be impossible to satisfy all the necessary boundary conditions, for the order of equation (C6) is lower than that equation (C1). (Reasoning along similar lines, it follows from the equations of hydrodynamics that necessary concomitants of geostrophic flow in rapidly rotating fluids are detached shear layers, a finding which has been amply confirmed by laboratory experiments and by the discovery of frontal systems in atmospheres and oceans (see Hide, R., 'Experiments with rotating fluids; Presidential Address', *Quart. J. R. Meteorol. Soc.* 103, 1-28 (1977)).

We conclude this appendix with inferences from electromagnetism concerning the generation of the MGF by the subtle self-exciting dynamo process, which occurs on timescales for which the right-hand side of equation (C1) cannot be neglected and involves the creation (and destruction) of field lines. The value of R must be high enough for efficient amplification of the strength of the field within the core by motional induction (as represented by the term curl ($\boldsymbol{u}$x$\boldsymbol{B}$)), which has to overcome ohmic decay (as represented by $(\mu\sigma)^{-1}$(divgrad)$\boldsymbol{B}$). But R must not be so high that the diffusion of the field from the core into the mantle is unduly inhibited, for the flux linkage of a perfect conductor cannot be changed.

In my own work I have found it useful to define dynamo action in terms of N at the CMB (rather than in terms of the magnetic energy or the equivalent magnetic moment of the system), requiring only that this quantity should not tend to zero as t tends to infinity (Hide, R., 'The magnetic flux linkage of a moving medium: a theorem and geophysical applications', *J. Geophys. Res.* 86, 11,681-11,687 (1981)). With this definition it is possible to express criteria for dynamo action in terms of the structure of $\boldsymbol{B}$ just below the CMB. It is also possible to show that irrespective of the compressibility of the fluid no steady or fluctuating magnetic field can be maintained by motional induction if the field possesses an axis of symmetry (R. Hide & T.N. Palmer, 'Generalization of Cowling's theorem', *Geophys. Astrophys. Fluid Dyn.* 19, 301-309, (1982)). This is an extension of a '*non-existence*' theorem due to T.G. Cowling [10], published in 1934 as an attempted rebuttal of J. Larmor's original suggestion [22] that sunspot magnetic fields are produced and maintained by dynamo action [10]. Not until the late 1950s did the first *existence* theorems (for classes of magnetic fields that do *not* possess an axis of symmetry), by Backus and Herzenberg [8-10], make their appearance in the literature. According to these theorems, there must exist (for values of R that are neither too small nor too large) non-decaying configurations of $\boldsymbol{u}$ and $\boldsymbol{B}$ that satisfy equation (C1). Finding and investigating configurations that are also consistent with the laws of mechanics and thermodynamics is the principal aim of dynamo theory [10].

TABLES

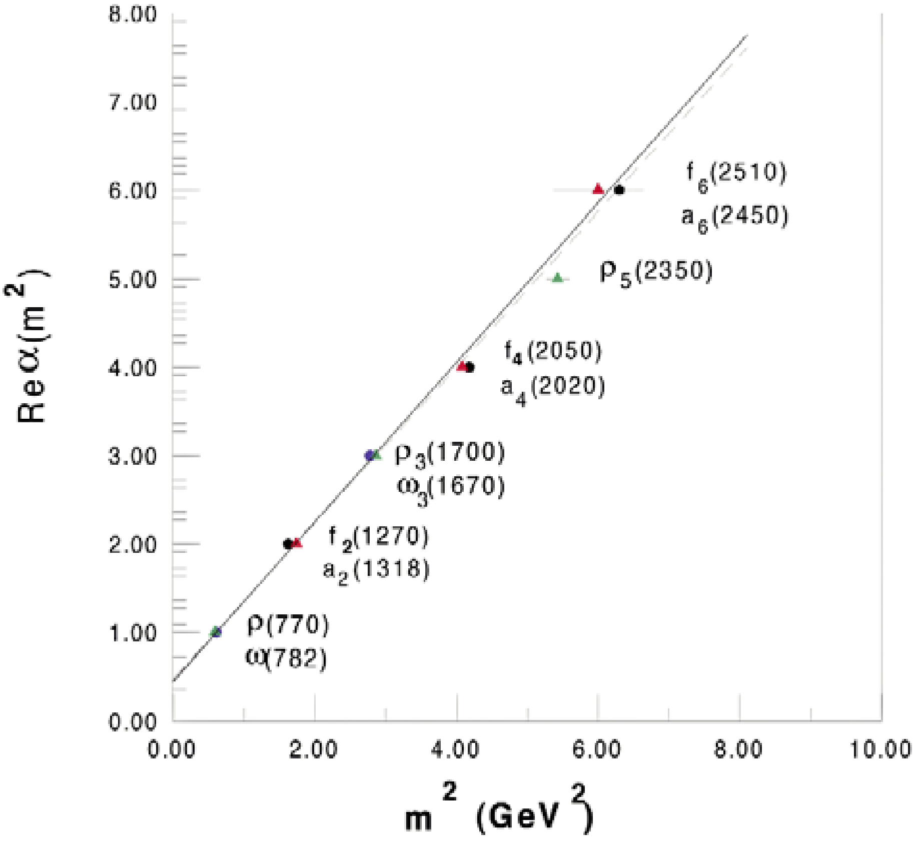


Figure 1. Chew-Frautschi plot (J versus m^2) represents almost perfect fit for observed hadrons (mesons) spin-mass relation. Taken from [3].

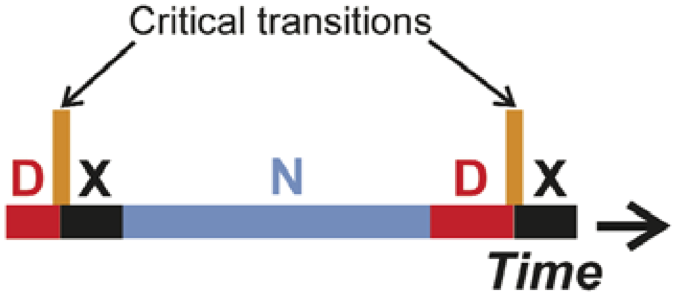
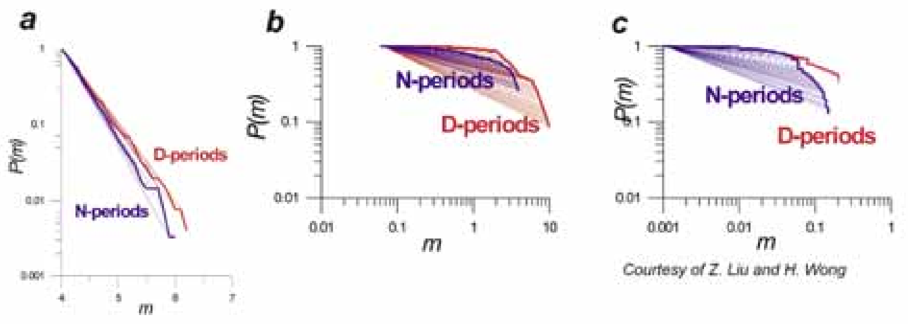
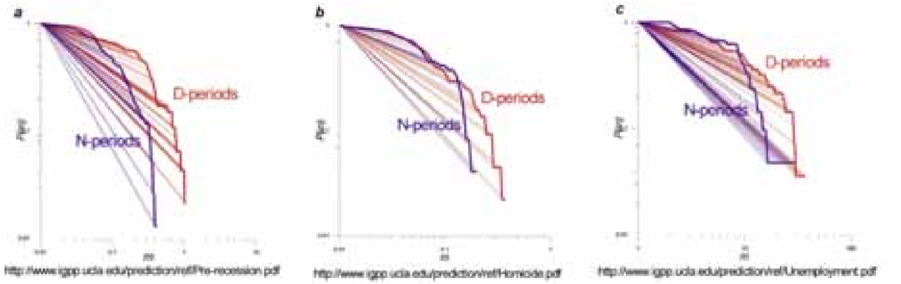


Figure 1. Division of time into periods of three kinds by their relation to critical transitions.



Courtesy of Z. Liu and H. Wong

Figure 2. Change of scaling before strong earthquakes. Southern California, 1954-2006 (*a*), California, 1968-2005 (*b, c*).



http://www.igpp.ucla.edu/prediction/rref/Pre-recession.pdf

http://www.igpp.ucla.edu/prediction/rref/Homicide.pdf

http://www.igpp.ucla.edu/prediction/rref/Unemployment.pdf

Figure 3. Change of scaling before socio-economic crises. Economic recessions, US, 1961-2002 (*a*); Surges of unemployment, US, 1961-2005 (*b*), Surge of homicides, Los Angeles, 1975-1993 (*c*).

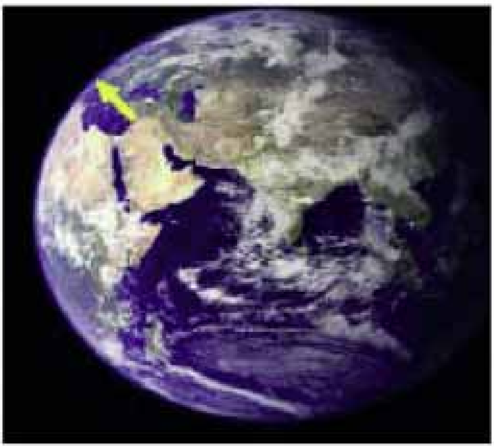


Figure 1.

Figure 2.

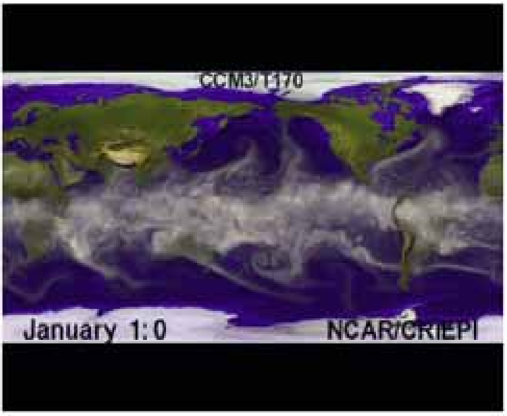


Figure 3. Ref. W.M. Washington.

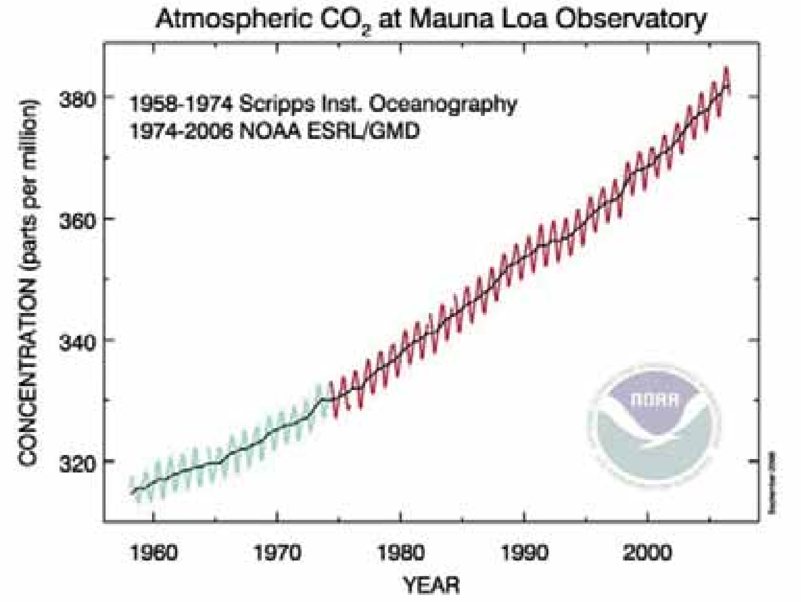


Figure 4. The Keeling Curve.

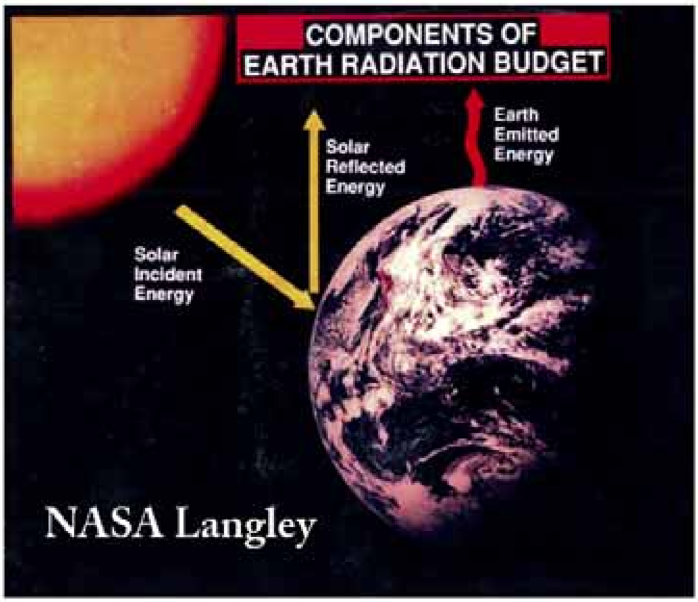


Figure 5. Earth Radiation Budget.

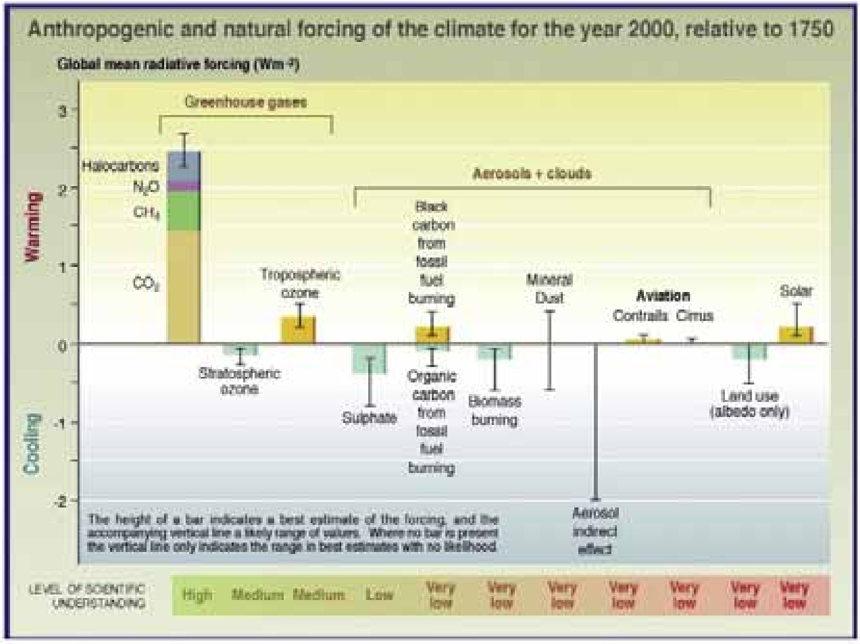


Figure 10. IPCC-AR3, 2001.

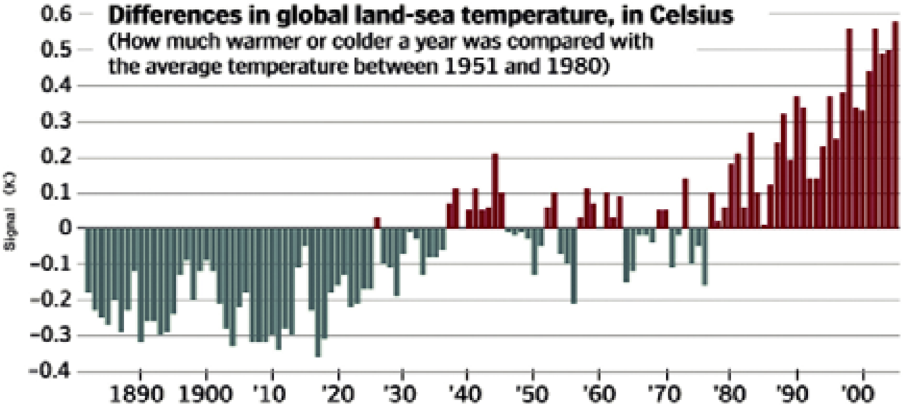
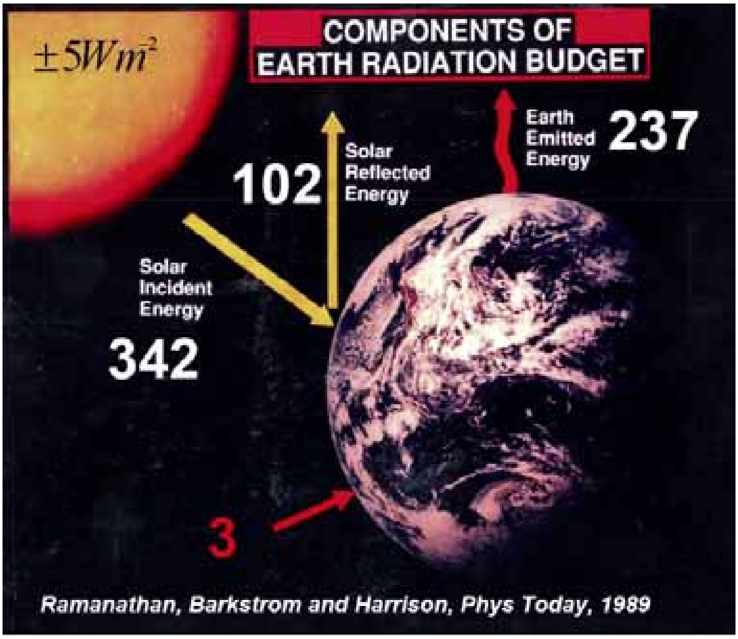


Figure 12. GISS Temperature Record. NASA_GISS.

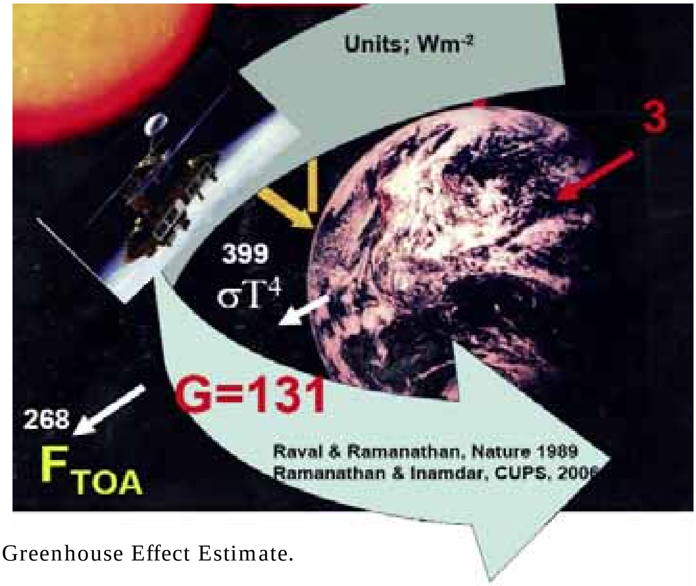
Figure 13. ERBE Satellite.



COMPONENTS OF EARTH RADIATION BUDGET

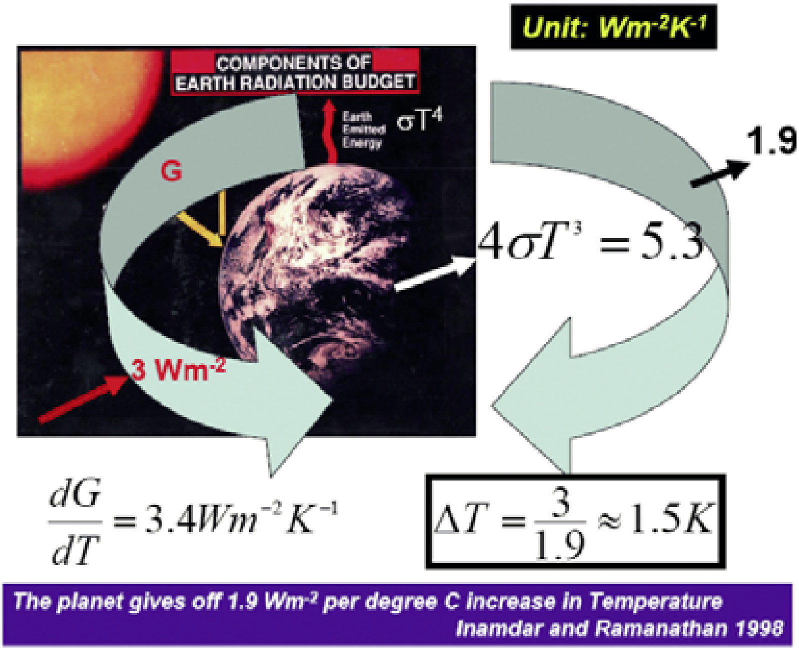
Ramanathan, Barkstrom and Harrison, Phys Today, 1989

Figure 14. ERBE Data.



Raval & Ramanathan, Nature 1989
Ramanathan & Inamdar, CUPS, 2006

Figure 15. Greenhouse Effect Estimate.



$$4\sigma T^3 = 5.3$$

$$\frac{dG}{dT} = 3.4Wm^{-2}K^{-1}$$

$$\boxed{\Delta T = \frac{3}{1.9} \approx 1.5K}$$

The planet gives off 1.9 Wm-2 per degree C increase in Temperature
Inamdar and Ramanathan 1998

Figure 18. Climate Sensitivity.

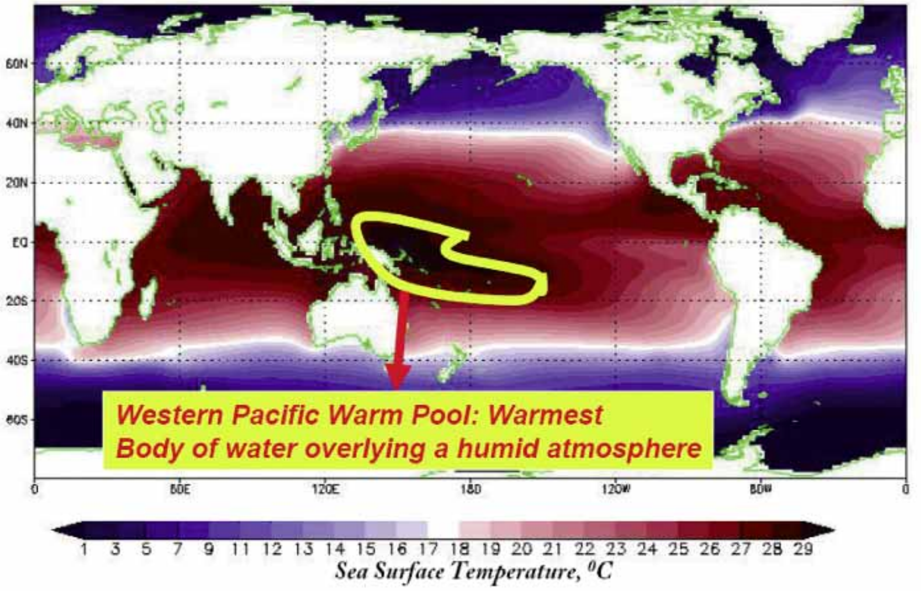


Figure 19. Observed Sea Surface Temp.

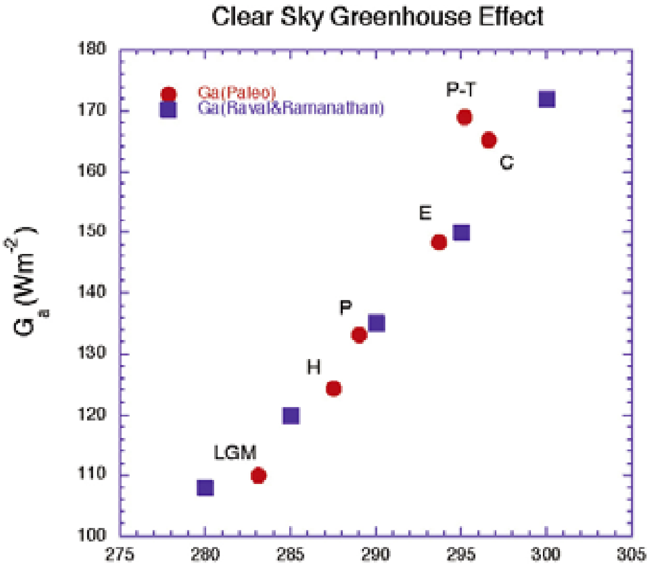


Figure 20. Kiehl, 2004. Private Comm.

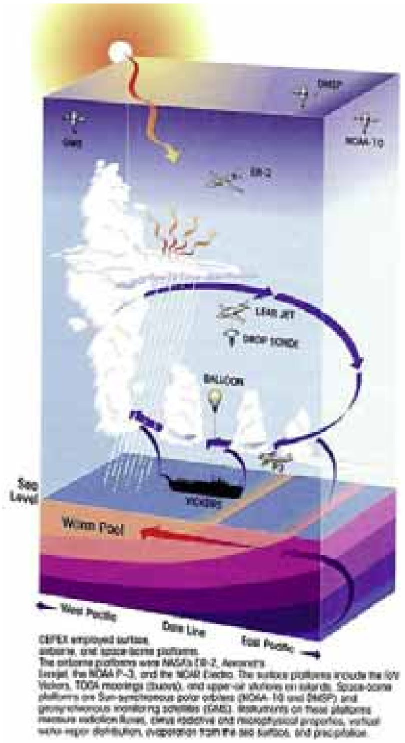

DMSP
GMS
NOAA-10
ER-2
LEAR JET
DROP SONDE
BALLOON
P3
Sea Level
VICKERS
Warm Pool
West Pacific
Date Line
East Pacific
CEPEX employed surface, airborne, and space-borne platforms. The airborne platforms were NASA's ER-2, Aeromet's Learjet, the NOAA P-3, and the NCAR Electra. The surface platforms include the R/V Vickers, TOGA moorings (buoys), and upper-air stations on islands. Space-borne platforms are Sun-synchronous polar orbiters (NOAA-10 and DMSP) and geosynchronous monitoring satellites (GMS). Instruments on these platforms measure radiation fluxes, cirrus radiative and microphysical properties, vertical water vapor distribution, evaporation from the sea surface, and precipitation.

Figure 21. CEPEX Configuration.

Figure 22. ABC Mask.

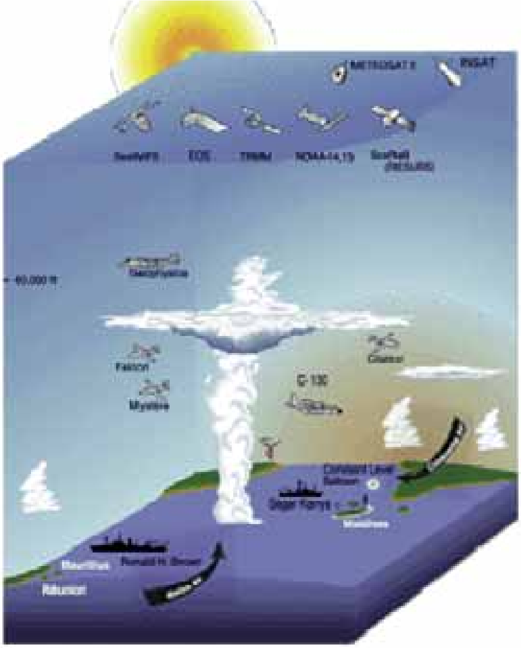


Figure 23. INDOEX Configuration.

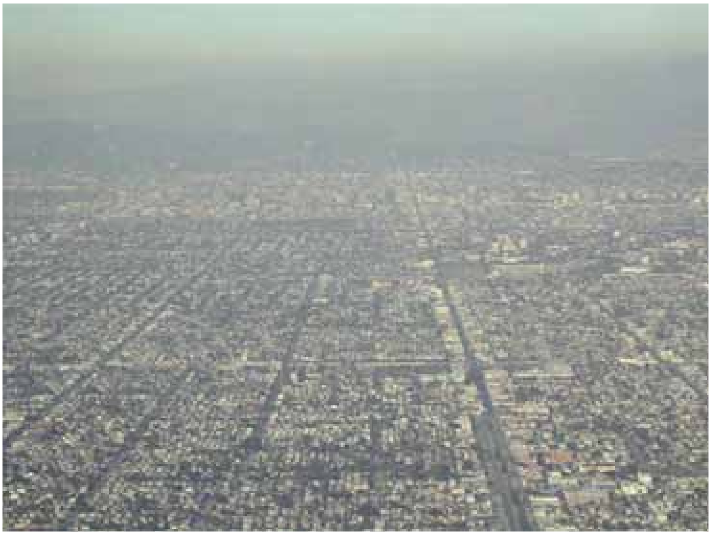


Figure 24. Los Angeles, Dec. 27 2002.

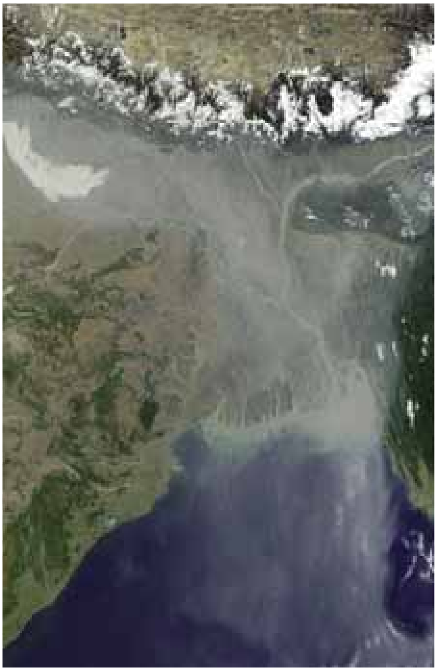


Figure 25. ABCs over India.

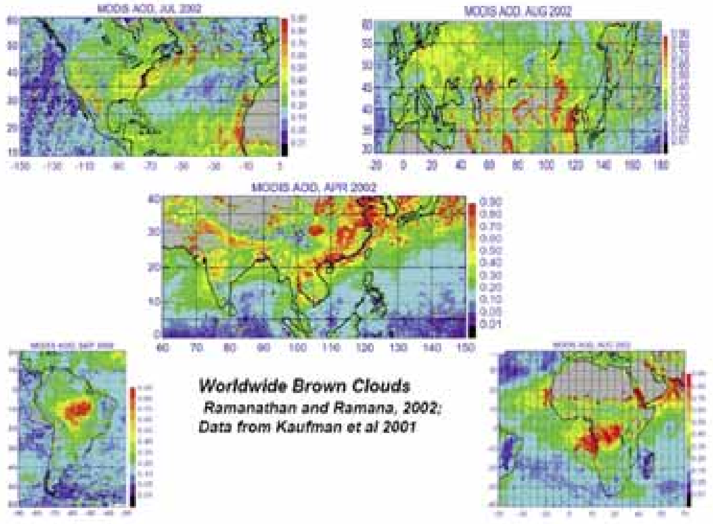
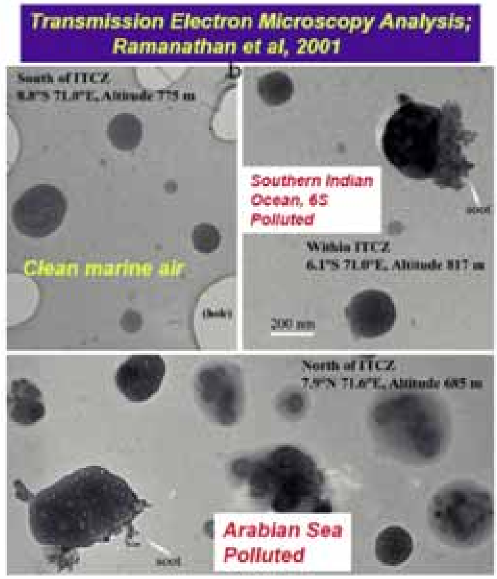


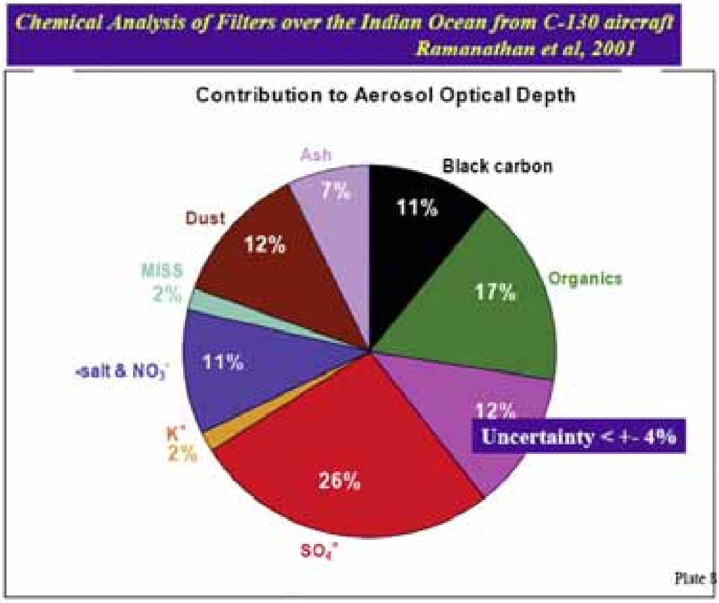
Figure 26. Satellite Aerosol Optical Depth.



Transmission Electron Microscopy Analysis; Ramanathan et al, 2001

b
South of ITCZ
8.8°S 71.0°E, Altitude 775 m
Clean marine air
(hole)
Southern Indian Ocean, 6S Polluted
soot
Within ITCZ
6.1°S 71.0°E, Altitude 817 m
200 nm
North of ITCZ
7.9°N 71.6°E, Altitude 685 m
Arabian Sea Polluted
soot

Figure 27. Filters from Indian Ocean.



Chemical Analysis of Filters over the Indian Ocean from C-130 aircraft Ramanathan et al, 2001

Figure 28. Chemical Speciation.

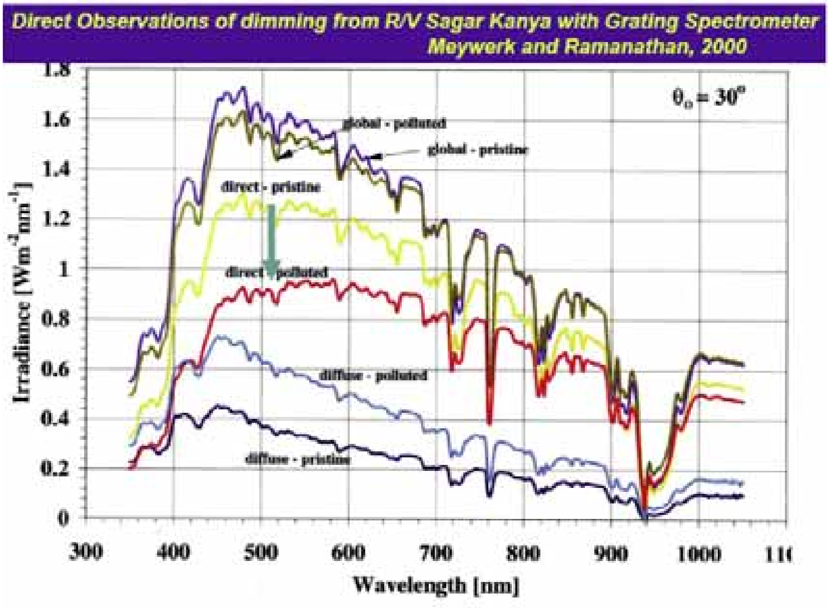


Figure 29.

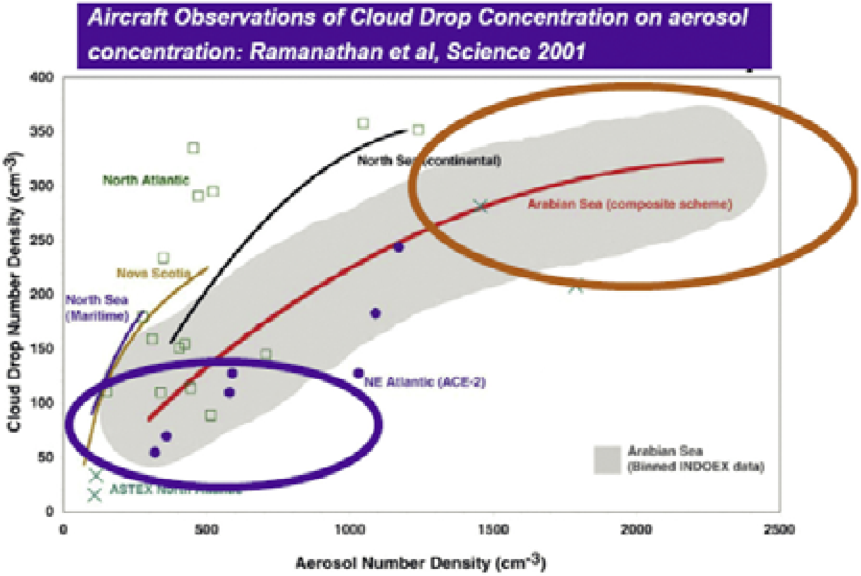


Figure 30.

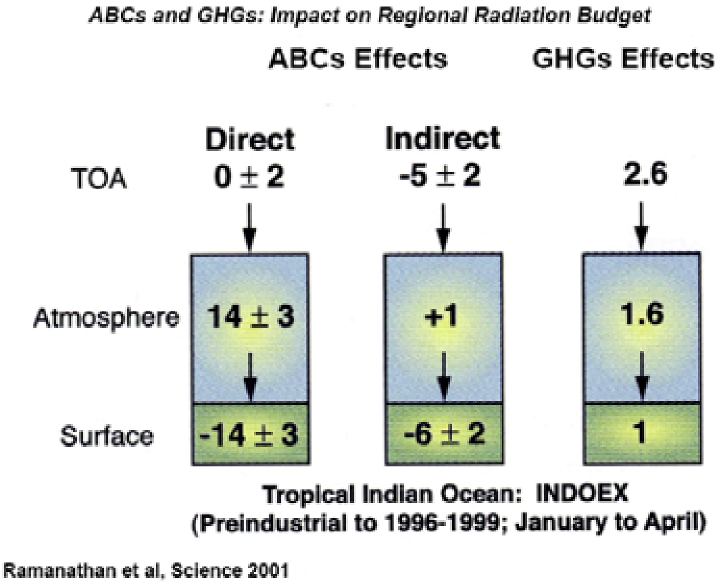

ABCs and GHGs: Impact on Regional Radiation Budget
ABCs Effects
GHGs Effects
Direct
Indirect
TOA
0 ± 2
-5 ± 2
2.6
Atmosphere
14 ± 3
+1
1.6
Surface
-14 ± 3
-6 ± 2
1
Tropical Indian Ocean: INDOEX
(Preindustrial to 1996-1999; January to April)
Ramanathan et al, Science 2001

Figure 31.

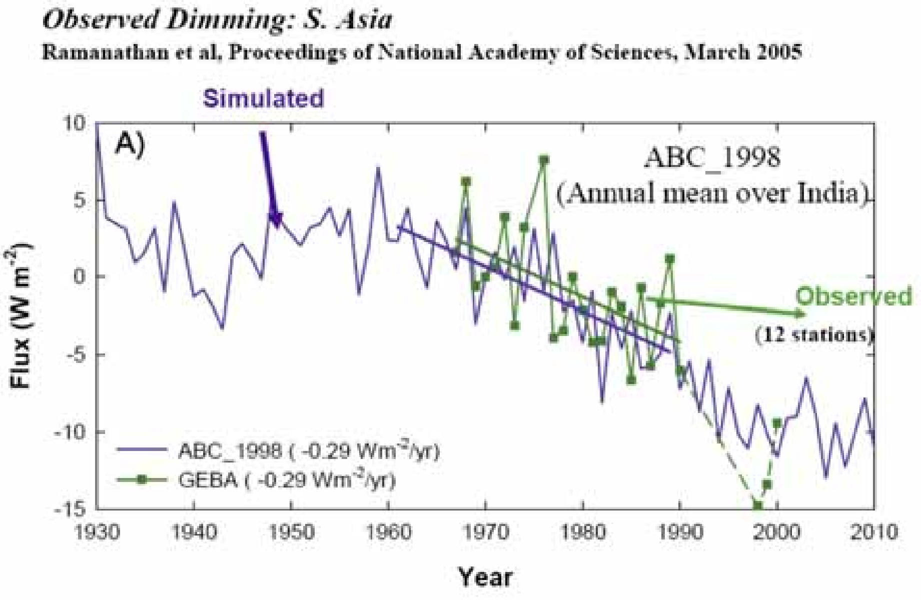


Figure 32.

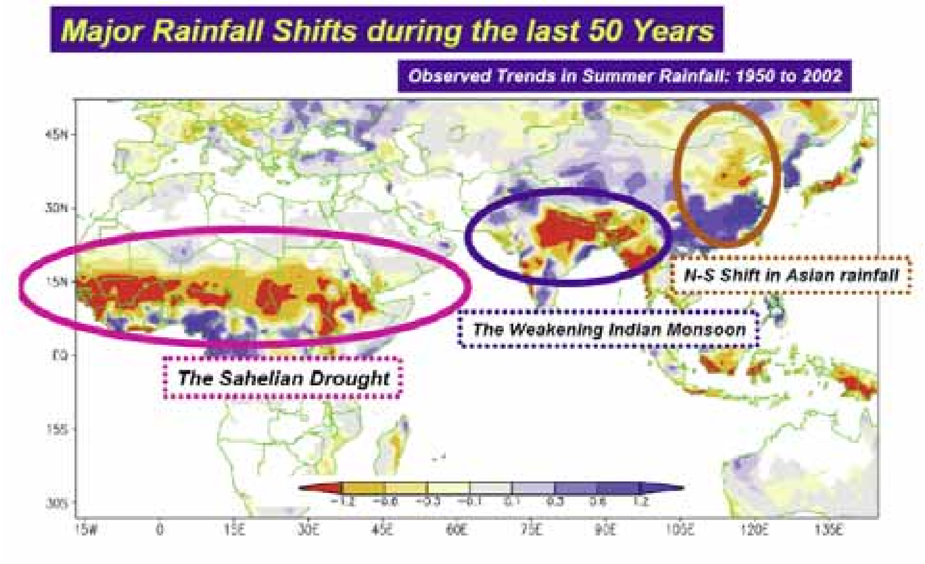
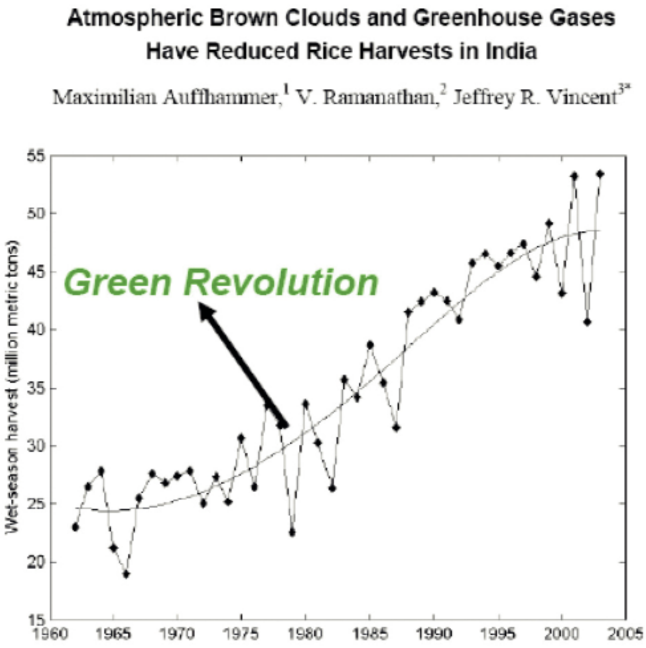
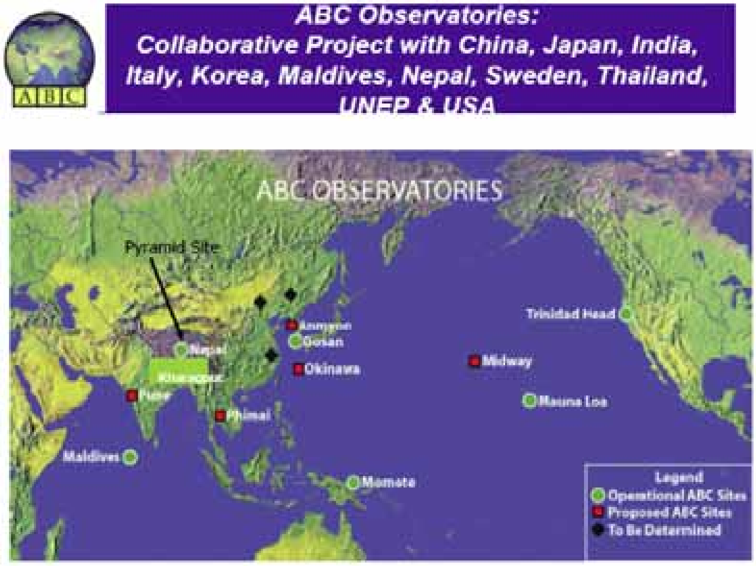

Major Rainfall Shifts during the last 50 Years
Observed Trends in Summer Rainfall: 1950 to 2002
N-S Shift in Asian rainfall
The Weakening Indian Monsoon
The Sahelian Drought
45N
30N
15N
EQ
15S
30S
15W
0
15E
30E
45E
60E
75E
90E
105E
120E
135E
-1.2
-0.6
-0.3
-0.1
0.1
0.3
0.6
1.2

Figure 33.


Atmospheric Brown Clouds and Greenhouse Gases Have Reduced Rice Harvests in India

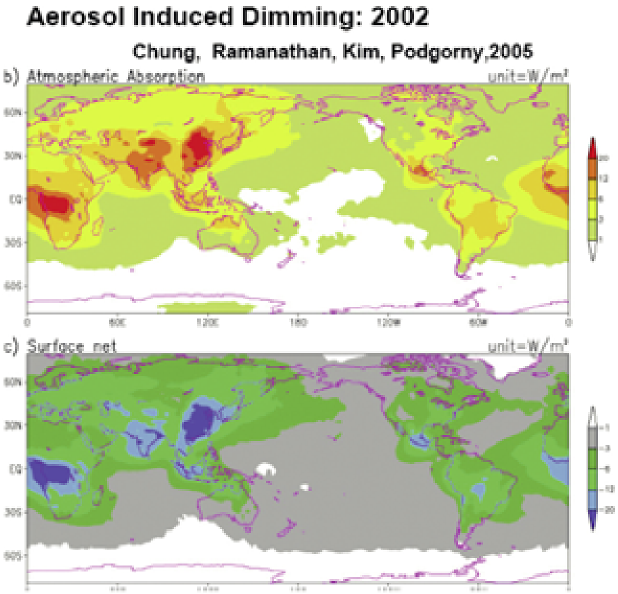
Maximilian Auffhammer,[1] V. Ramanathan,[2] Jeffrey R. Vincent[3*]

Figure 34.



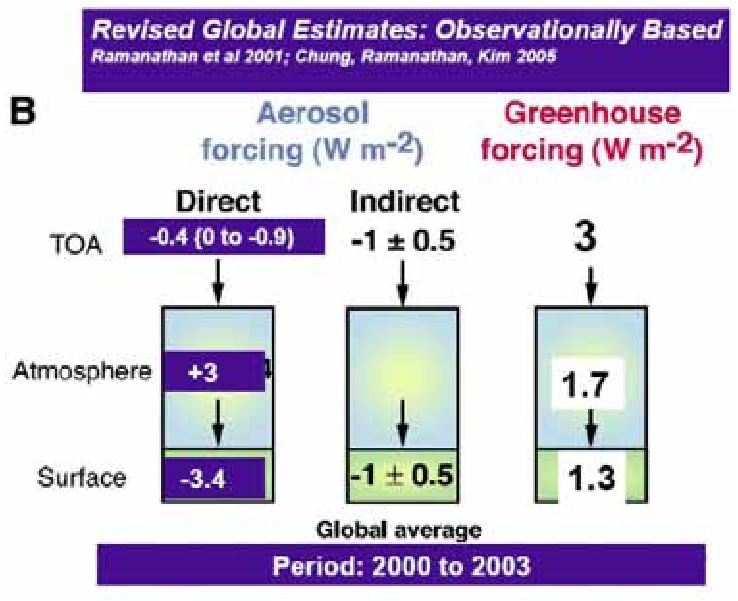
ABC Observatories:
Collaborative Project with China, Japan, India, Italy, Korea, Maldives, Nepal, Sweden, Thailand, UNEP & USA

Figure 35.



Aerosol Induced Dimming: 2002
Chung, Ramanathan, Kim, Podgorny,2005
b) Atmospheric Absorption
unit=W/m²
60N
30N
EQ
30S
60S
0
60E
120E
180
120W
60W
0
20
12
6
3
1
c) Surface net
unit=W/m²
-1
-3
-6
-12
-20

Figure 36. ABC forcing; solar heating of the atmosphere (Top panel); Dimming (bottom panel).

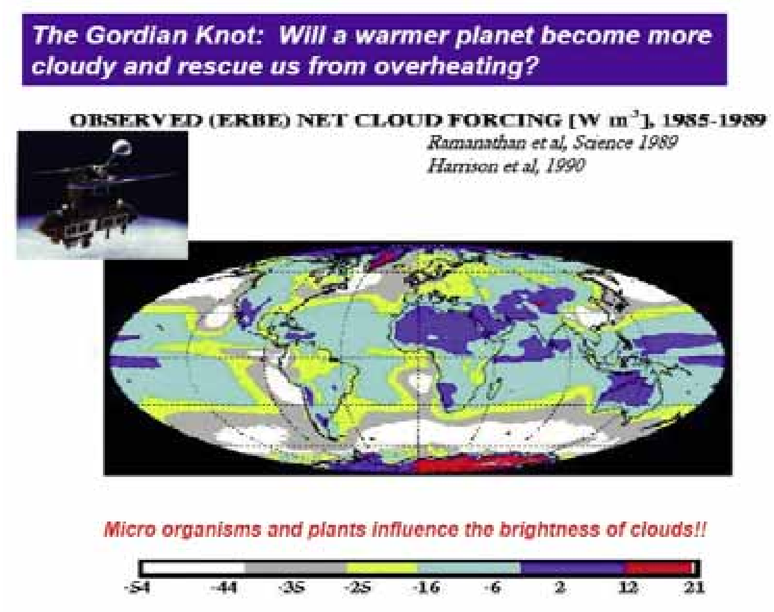


Revised Global Estimates: Observationally Based

Ramanathan et al 2001; Chung, Ramanathan, Kim 2005

Period: 2000 to 2003

Figure 37.



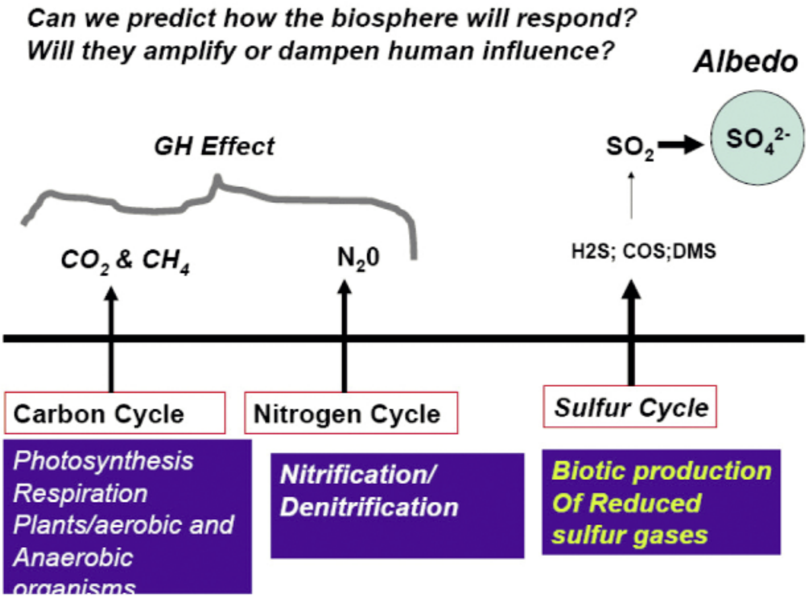
The Gordian Knot: Will a warmer planet become more cloudy and rescue us from overheating?

OBSERVED (ERBE) NET CLOUD FORCING [W m^{-2}], 1985-1989

Ramanathan et al, Science 1989
Harrison et al, 1990

Micro organisms and plants influence the brightness of clouds!!

Figure 38.



Can we predict how the biosphere will respond?
Will they amplify or dampen human influence?

Carbon Cycle

Photosynthesis
Respiration
Plants/aerobic and
Anaerobic
organisms

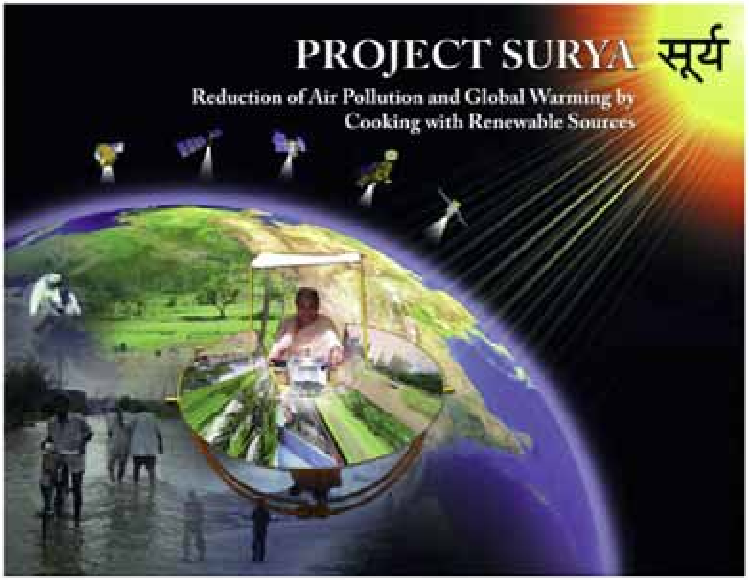
Nitrogen Cycle

Nitrification/
Denitrification

Sulfur Cycle

Biotic production
Of Reduced
sulfur gases

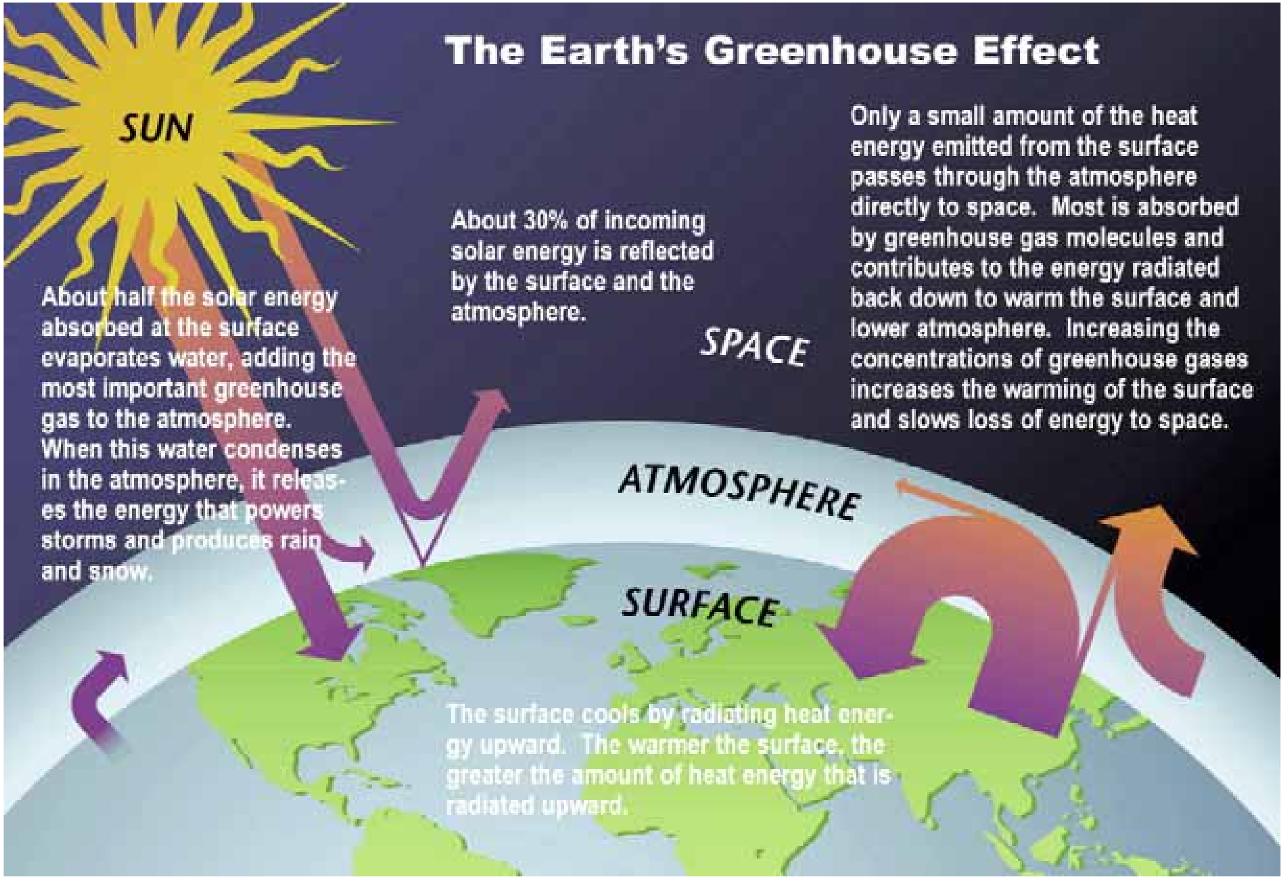
Figure 39.



PROJECT SURYA सूर्य

Reduction of Air Pollution and Global Warming by Cooking with Renewable Sources

Figure 40.



The Earth's Greenhouse Effect

About half the solar energy absorbed at the surface evaporates water, adding the most important greenhouse gas to the atmosphere. When this water condenses in the atmosphere, it releases the energy that powers storms and produces rain and snow.

About 30% of incoming solar energy is reflected by the surface and the atmosphere.

Only a small amount of the heat energy emitted from the surface passes through the atmosphere directly to space. Most is absorbed by greenhouse gas molecules and contributes to the energy radiated back down to warm the surface and lower atmosphere. Increasing the concentrations of greenhouse gases increases the warming of the surface and slows loss of energy to space.

The surface cools by radiating heat energy upward. The warmer the surface, the greater the amount of heat energy that is radiated upward.

Figure 1.

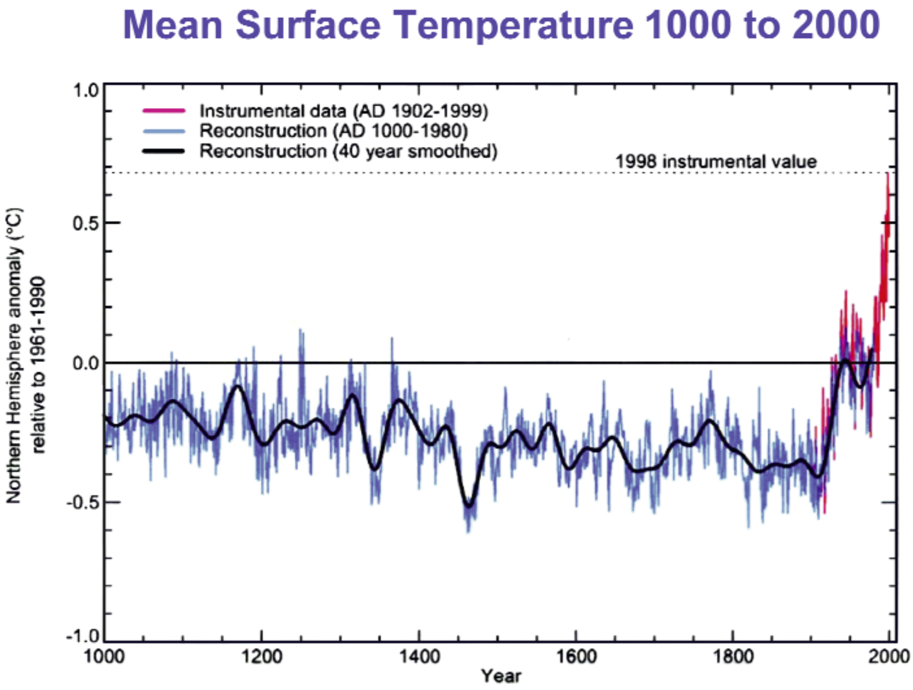
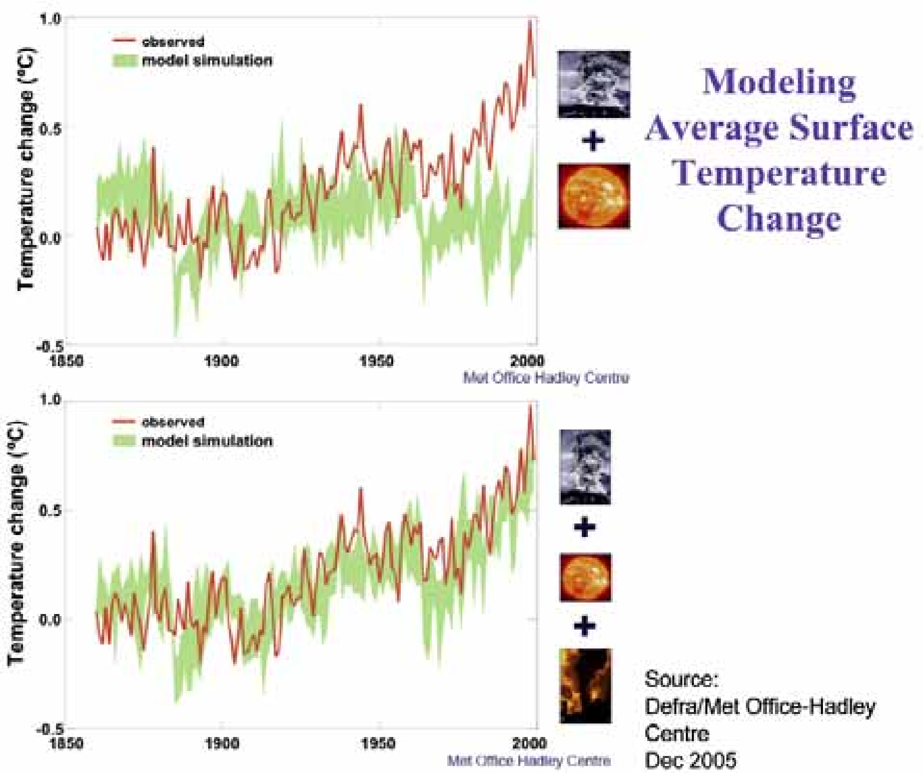

Mean Surface Temperature 1000 to 2000
Instrumental data (AD 1902-1999)
Reconstruction (AD 1000-1980)
Reconstruction (40 year smoothed)
1998 instrumental value
Northern Hemisphere anomaly (°C) relative to 1961-1990
1.0
0.5
0.0
-0.5
-1.0
1000
1200
1400
1600
1800
2000
Year

Figure 5.



Modeling Average Surface Temperature Change

Source:
Defra/Met Office-Hadley
Centre
Dec 2005

Figure 6.

Global Mean Surface Temperature 1880 - 2005

J. Hansen, R. Ruedy, M. Sato, and K. Lo

NASA Goddard Inst. for Space Studies
and Columbia Univ. Earth Institute

New York, NY 10025, USA

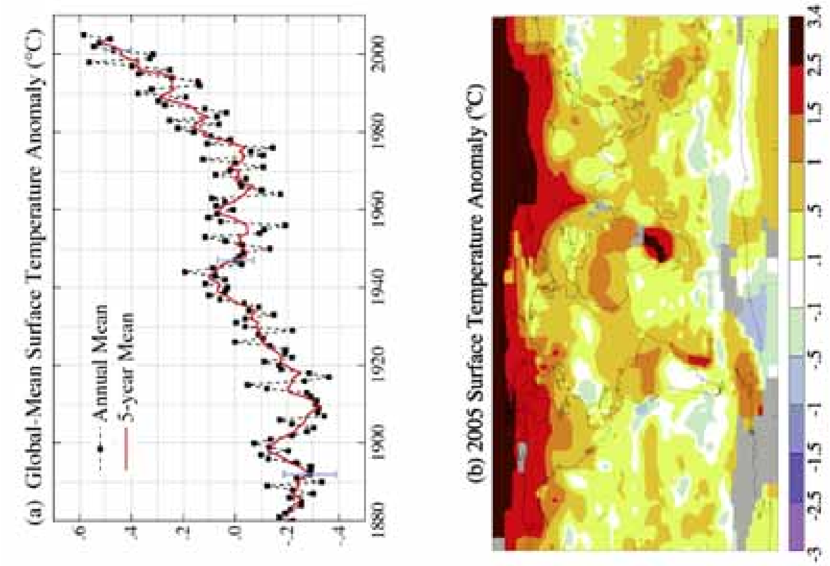


Figure 7.

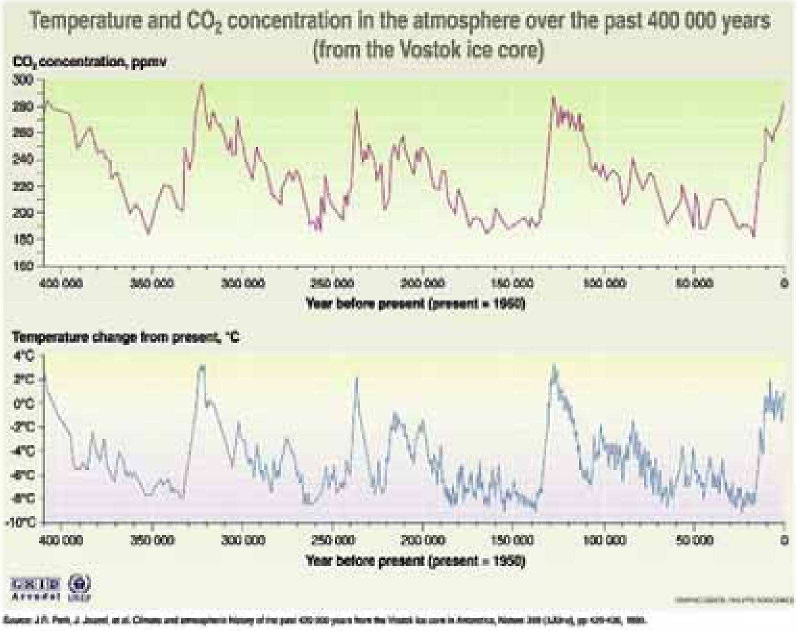


Figure 8.

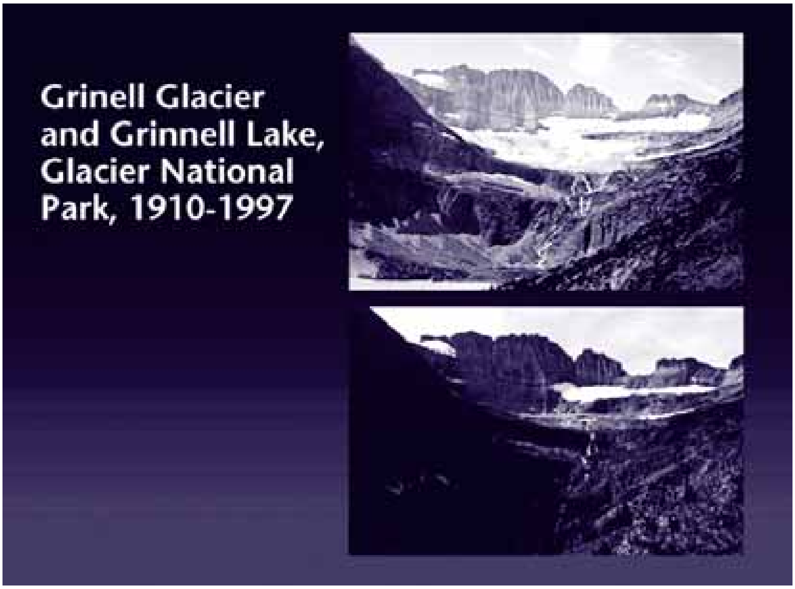
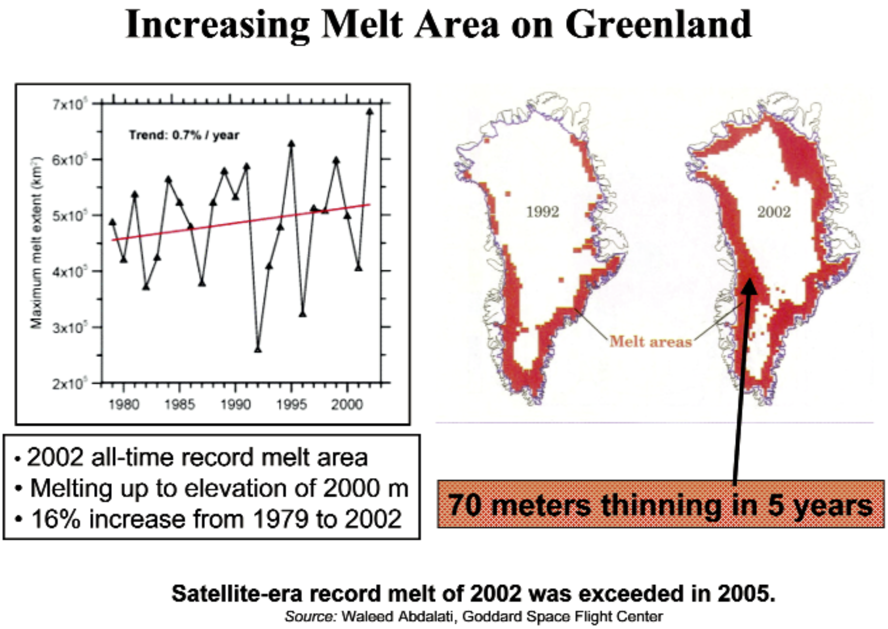

Grinell Glacier and Grinnell Lake, Glacier National Park, 1910-1997

Figure 9.



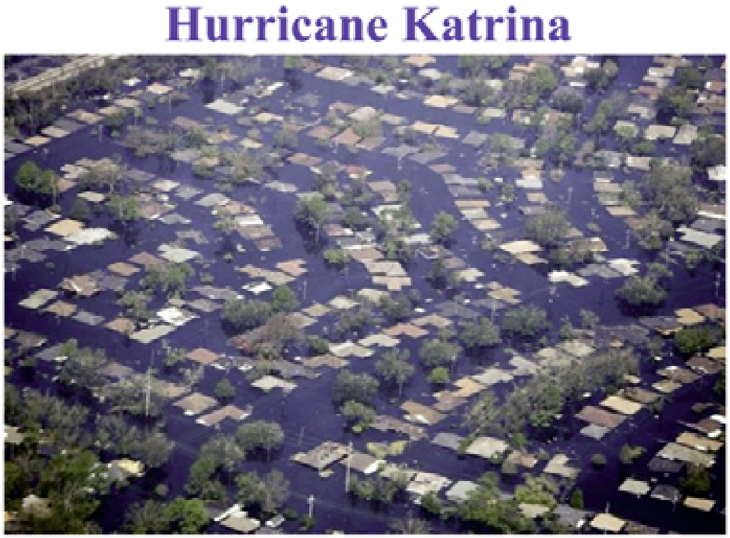
Increasing Melt Area on Greenland

- 2002 all-time record melt area
- Melting up to elevation of 2000 m
- 16% increase from 1979 to 2002

70 meters thinning in 5 years

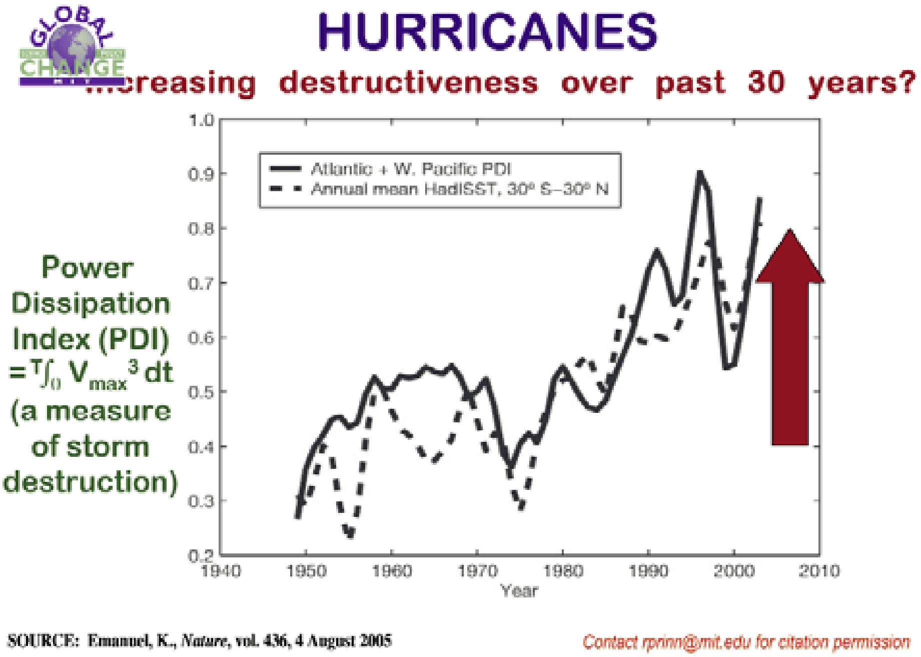
Satellite-era record melt of 2002 was exceeded in 2005.
Source: Waleed Abdalati, Goddard Space Flight Center

Figure 10.



Hurricane Katrina

Figure 11.



HURRICANES

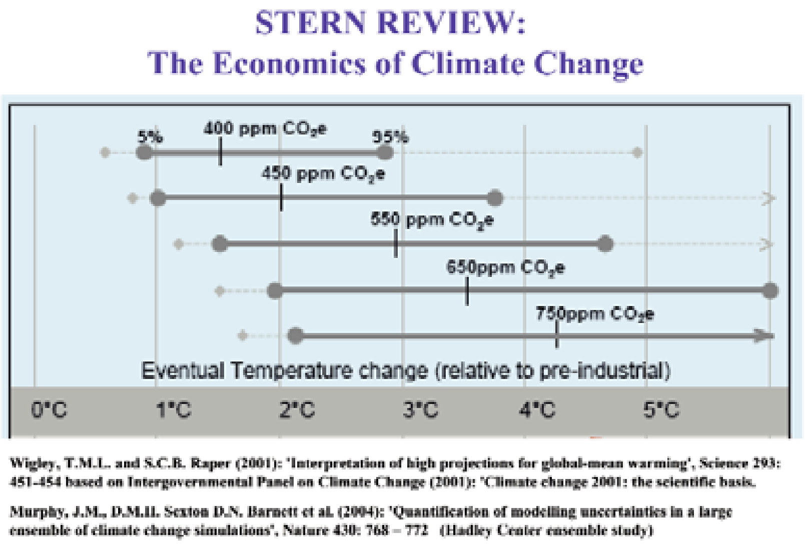
...creasing destructiveness over past 30 years?

Power Dissipation Index (PDI) $= {}^{T}\!\int_0 V_{max}{}^3\,dt$ (a measure of storm destruction)

SOURCE: Emanuel, K., *Nature*, vol. 436, 4 August 2005

Contact rprinn@mit.edu for citation permission

Figure 12.

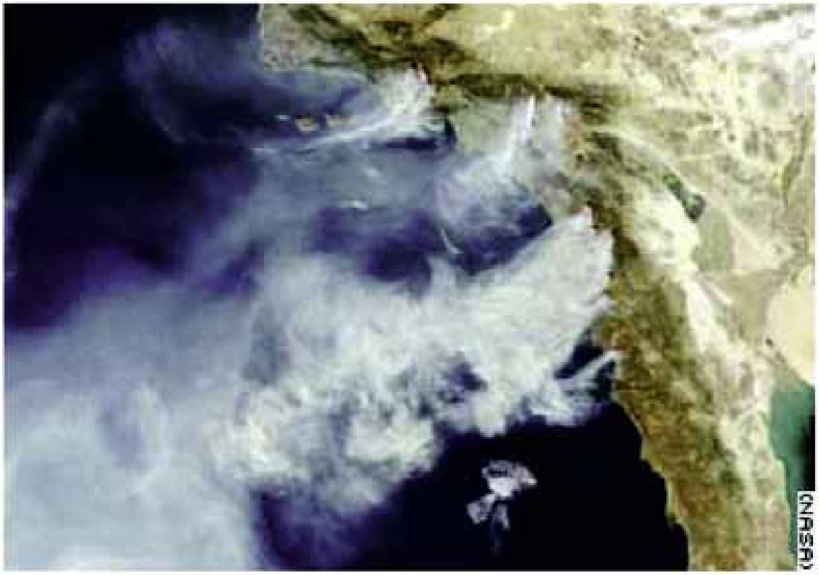


STERN REVIEW:
The Economics of Climate Change

Wigley, T.M.L. and S.C.B. Raper (2001): 'Interpretation of high projections for global-mean warming', Science 293: 451-454 based on Intergovernmental Panel on Climate Change (2001): 'Climate change 2001: the scientific basis.

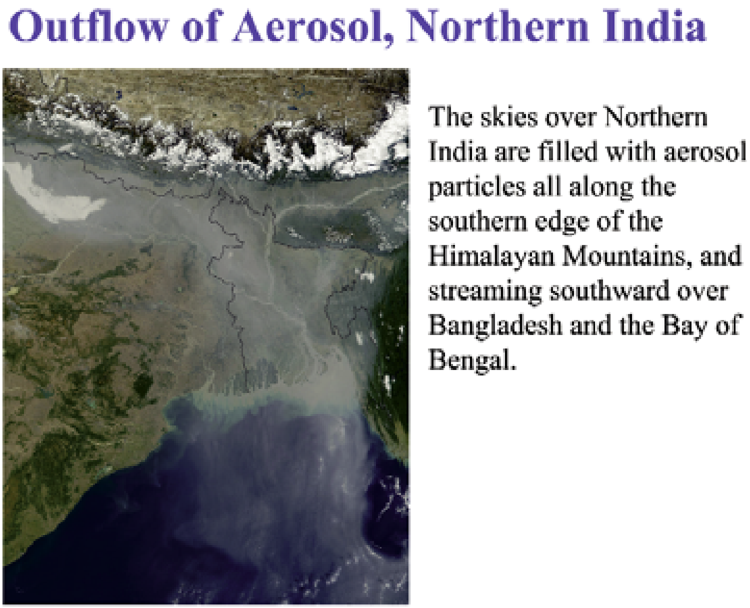
Murphy, J.M., D.M.H. Sexton D.N. Barnett et al. (2004): 'Quantification of modelling uncertainties in a large ensemble of climate change simulations', Nature 430: 768 – 772 (Hadley Center ensemble study)

Figure 13.



Satellite photo of smoke from S California wildfires, October 2003

Figure 14.



Outflow of Aerosol, Northern India

The skies over Northern India are filled with aerosol particles all along the southern edge of the Himalayan Mountains, and streaming southward over Bangladesh and the Bay of Bengal.

Figure 15.

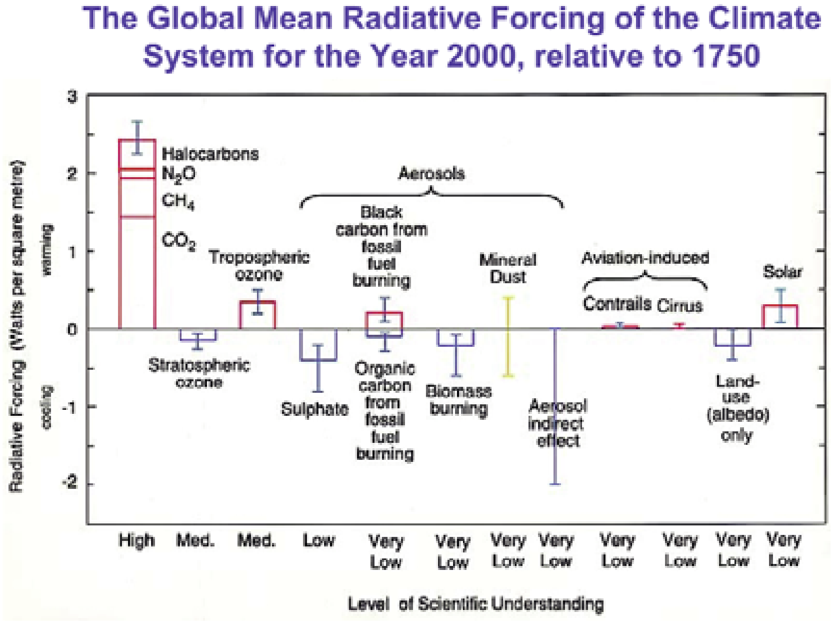
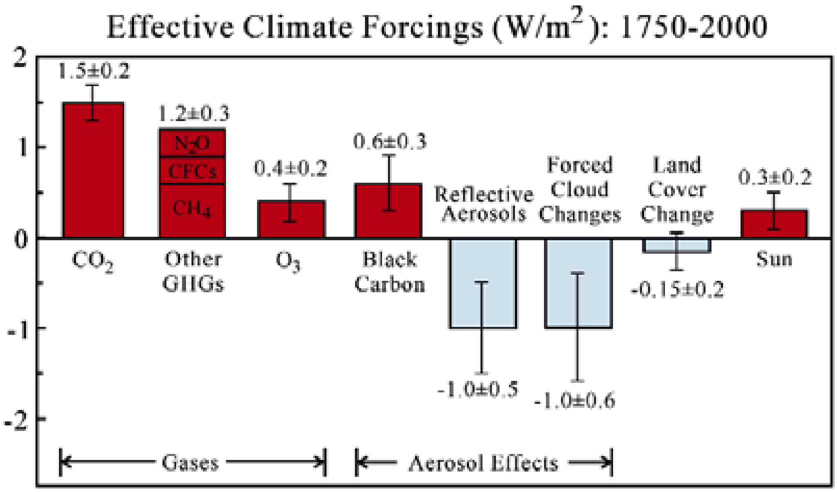


Figure 16.



Climate forcing agents in the industrial era. "Effective" forcing accounts for "efficacy" of the forcing mechanism

Source: Hansen et al., JGR, **110**, D18104, 2005.

Figure 17.

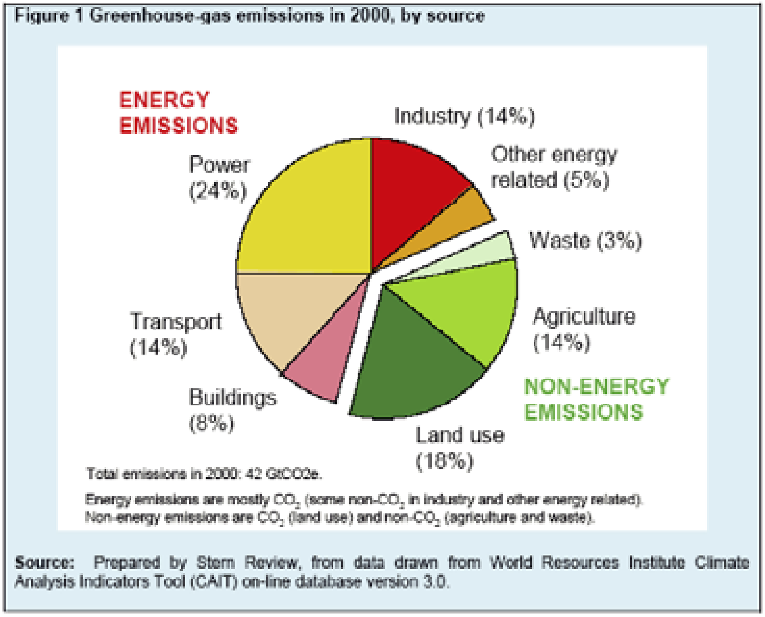


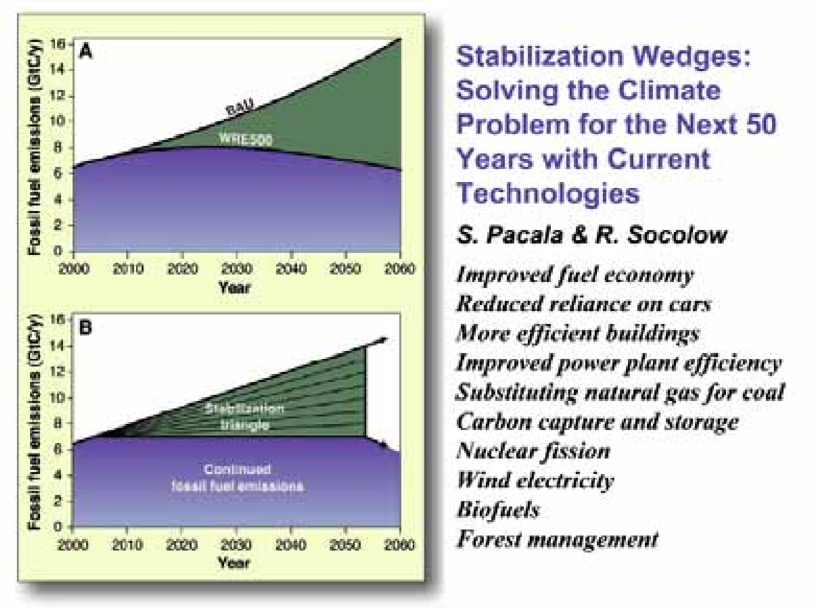
Figure 1 Greenhouse-gas emissions in 2000, by source

Total emissions in 2000: 42 GtCO2e.

Energy emissions are mostly CO_2 (some non-CO_2 in industry and other energy related). Non-energy emissions are CO_2 (land use) and non-CO_2 (agriculture and waste).

Source: Prepared by Stern Review, from data drawn from World Resources Institute Climate Analysis Indicators Tool (CAIT) on-line database version 3.0.

Figure 18.



Stabilization Wedges: Solving the Climate Problem for the Next 50 Years with Current Technologies

S. Pacala & R. Socolow

Improved fuel economy
Reduced reliance on cars
More efficient buildings
Improved power plant efficiency
Substituting natural gas for coal
Carbon capture and storage
Nuclear fission
Wind electricity
Biofuels
Forest management

Figure 19.

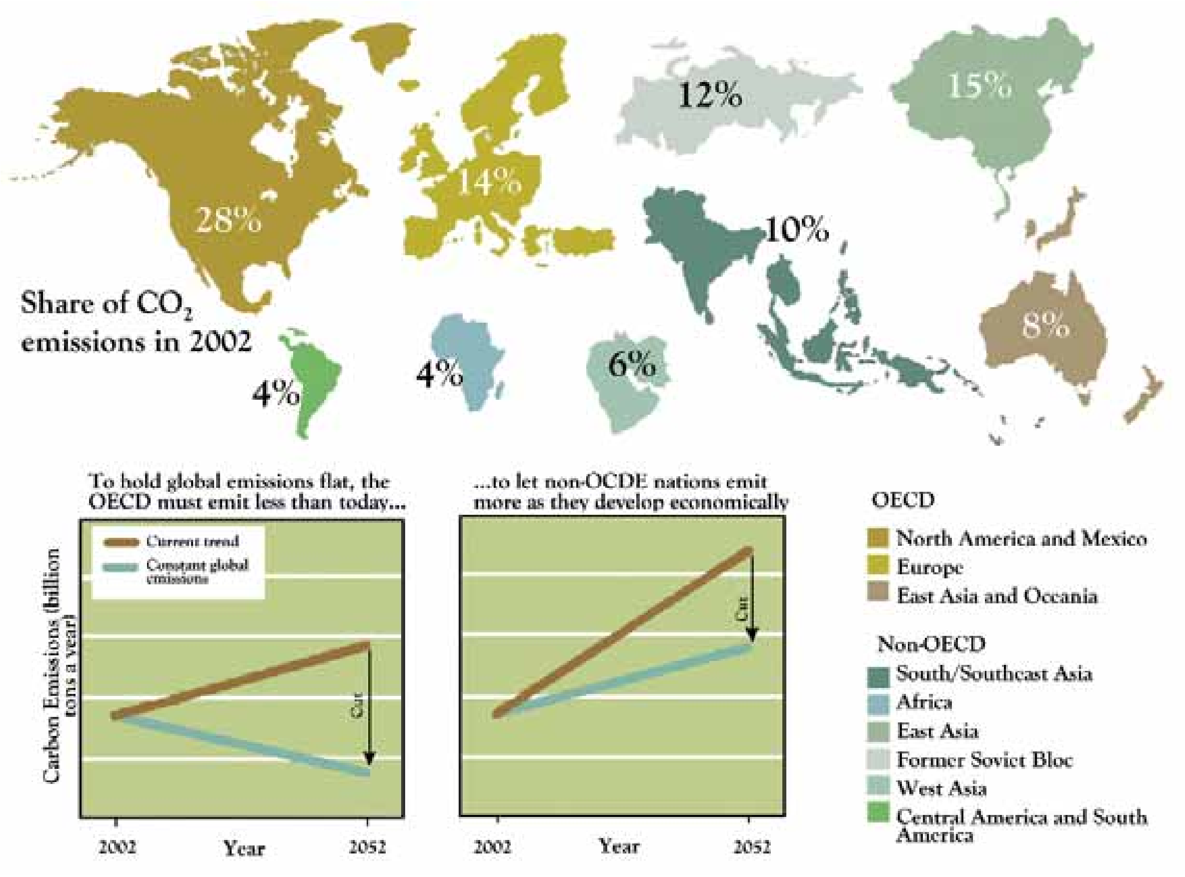
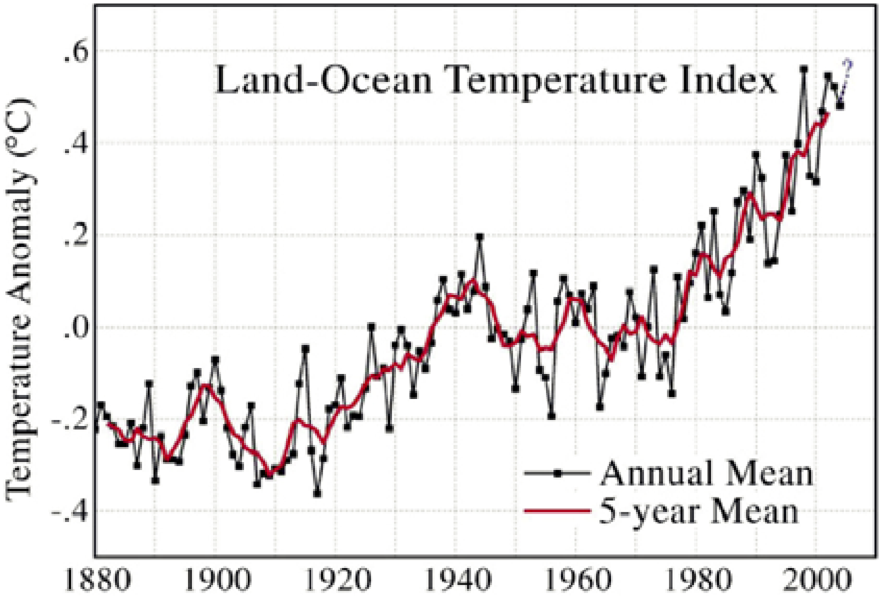

Share of CO_2 emissions in 2002
28%
14%
12%
15%
10%
8%
4%
4%
6%
To hold global emissions flat, the OECD must emit less than today...
Current trend
Constant global emissions
Carbon Emissions (billion tons a year)
Cut
2002
Year
2052
...to let non-OCDE nations emit more as they develop economically
Cut
2002
Year
2052
OECD
North America and Mexico
Europe
East Asia and Occania
Non-OECD
South/Southeast Asia
Africa
East Asia
Former Soviet Bloc
West Asia
Central America and South America

Figure 20.



Global mean surface temperature change based on surface air measurements over land and SSTs over ocean

Source: Update of Hansen et al., *JGR*, **106**, 23947, 2001; Reynolds and Smith, *J. Climate*, **7**, 1994; Rayner et al., *JGR*, **108**, 2003.

Figure 5.

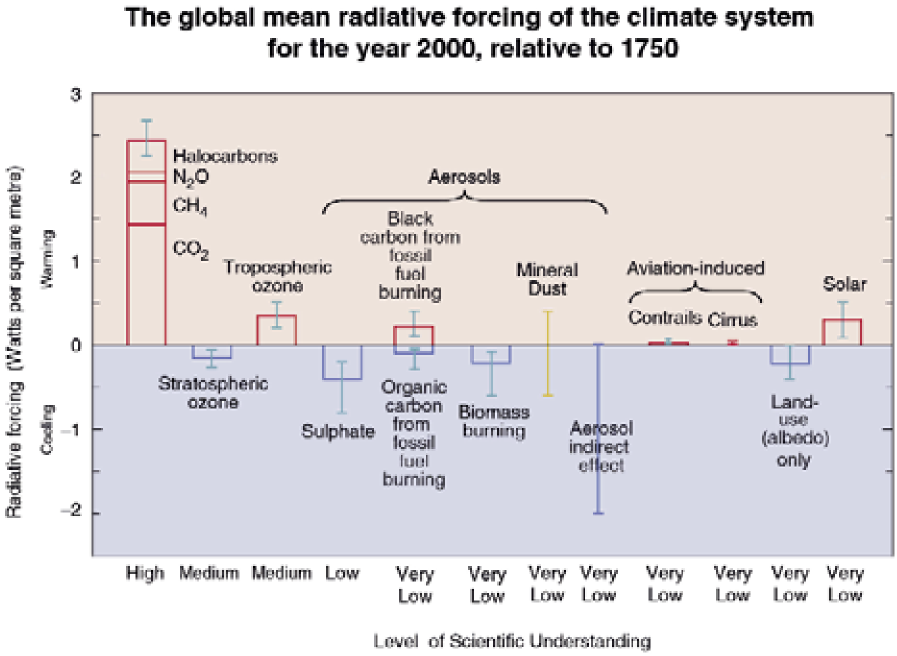


Figure 6.

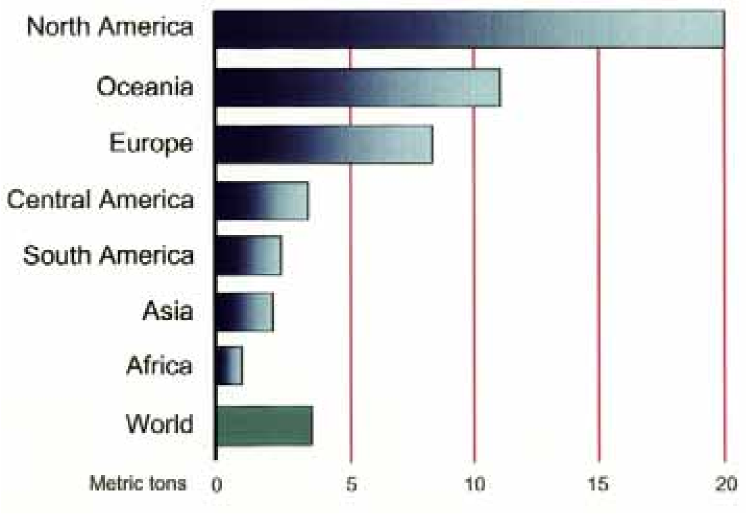
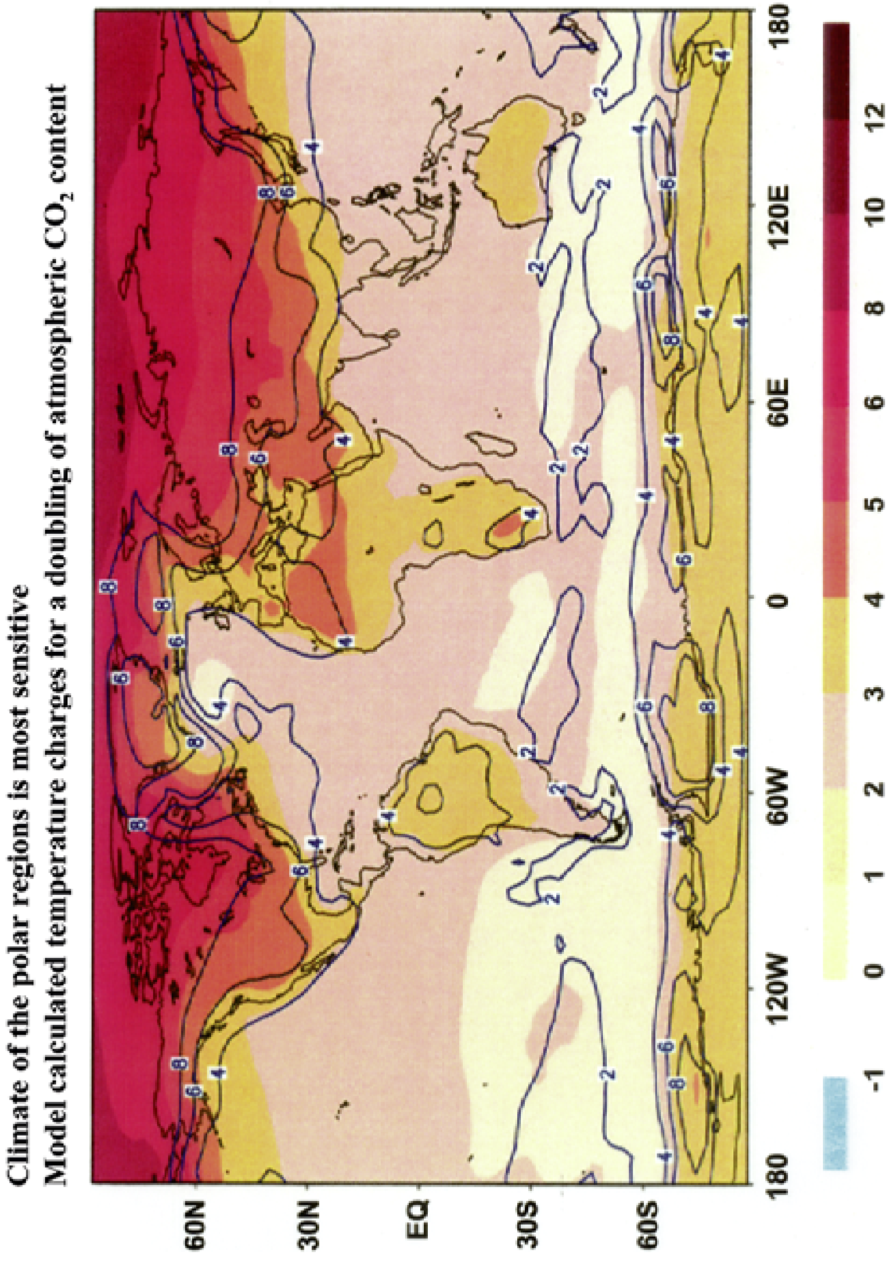


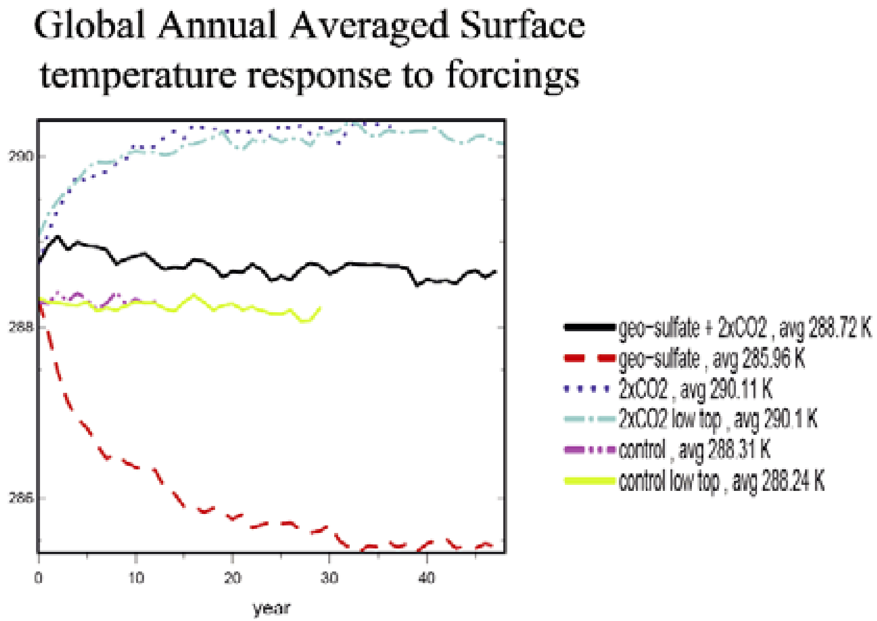
Figure 7.



Climate of the polar regions is most sensitive

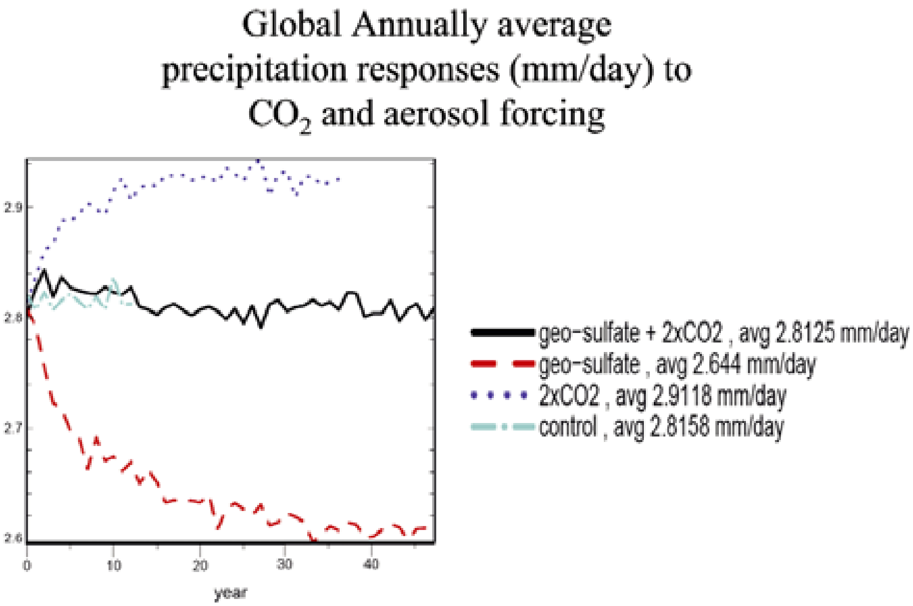
Model calculated temperature charges for a doubling of atmospheric CO_2 content

Figure 8.



Global Annual Averaged Surface temperature response to forcings

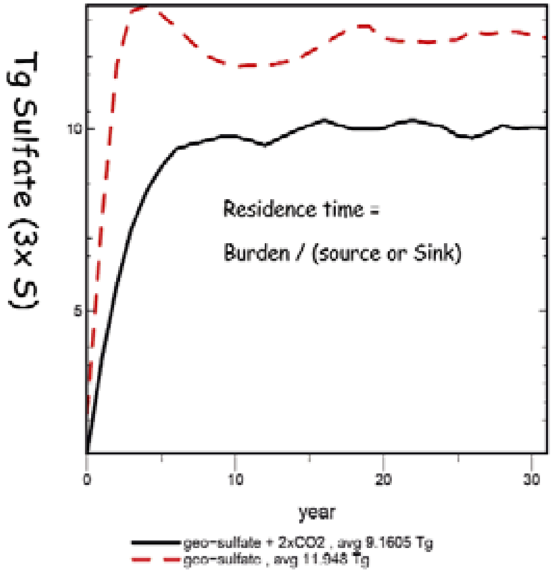
Figure 10.



Global Annually average precipitation responses (mm/day) to CO_2 and aerosol forcing

Figure 11.

Burden and Lifetime of Geo-Sulfate



Residence time ~ 4 years

Residence time ~ 3 years

Residence time longer than usually estimated from volcanic aerosol (1-2 years)

Residence time depends on State!
controlled by:

- cross tropopause exchange
- hydrologic cycle

Figure 12.

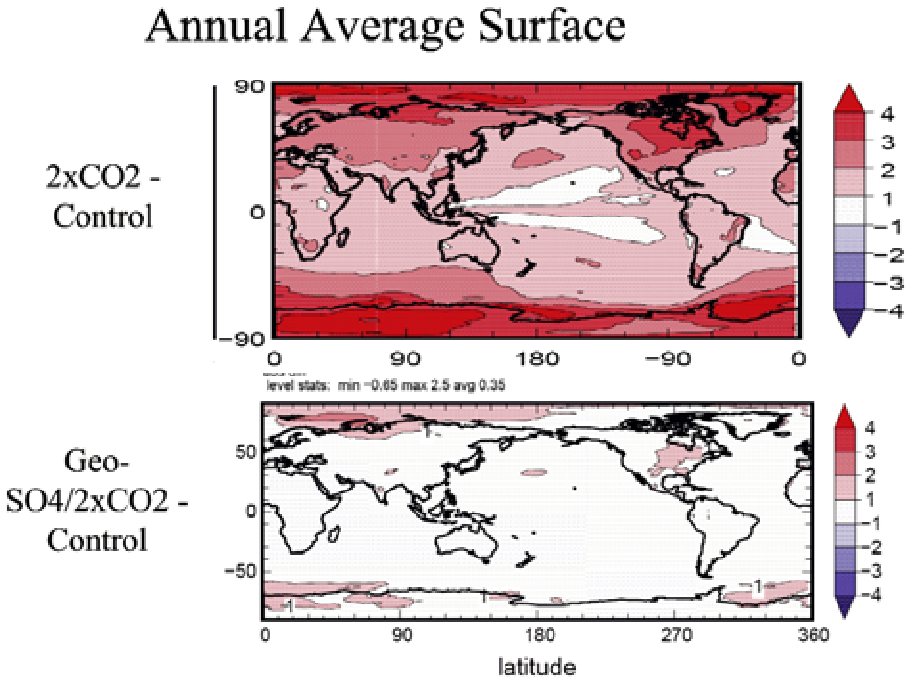
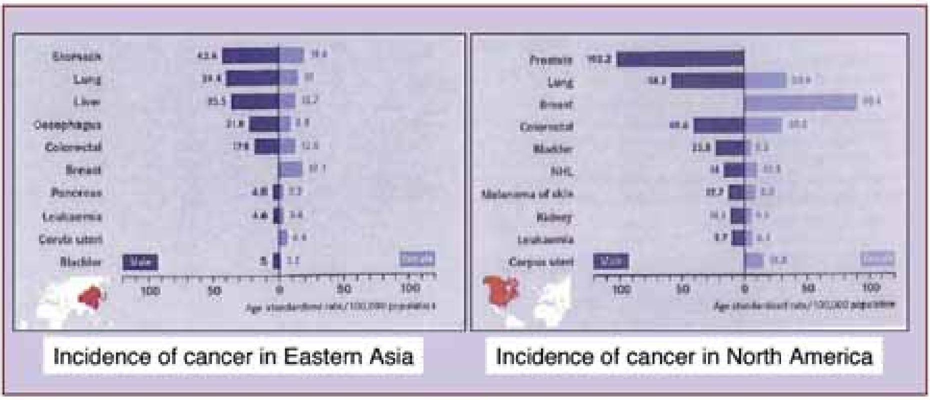

Annual Average Surface
2xCO2 -
Control
90
0
−90
0
90
180
−90
0
4
3
2
1
−1
−2
−3
−4
level stats: min −0.65 max 2.5 avg 0.35
Geo-
SO4/2xCO2 -
Control
50
0
-50
0
90
180
270
360
latitude
−1
1

Figure 13.



Incidence of cancer in Eastern Asia

Incidence of cancer in North America

Figure 3.

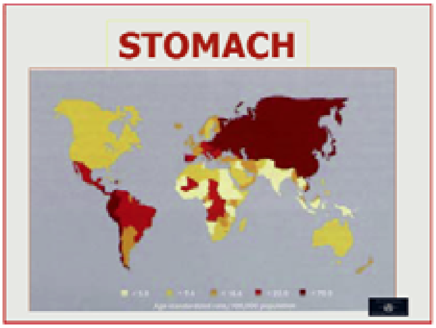

STOMACH

Figure 4.

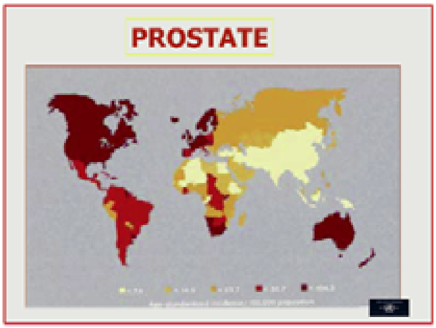

PROSTATE

Figure 5.

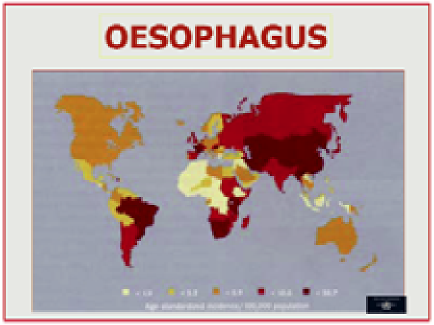

OESOPHAGUS

Figure 6.

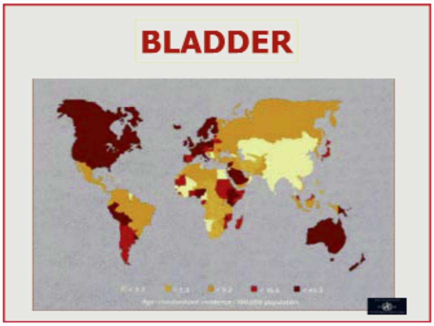

BLADDER

Figure 7.

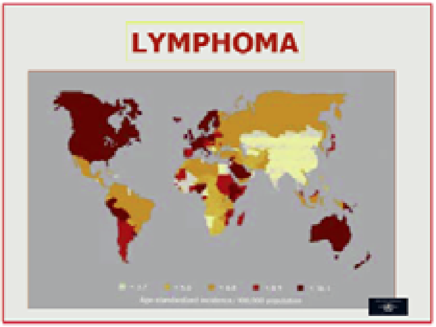


Figure 8.

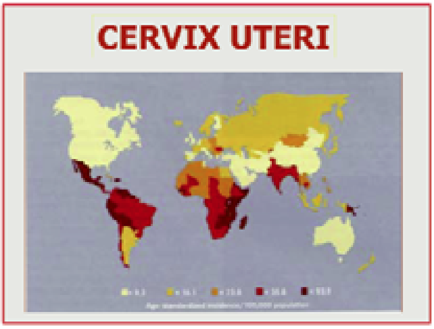


Figure 9.

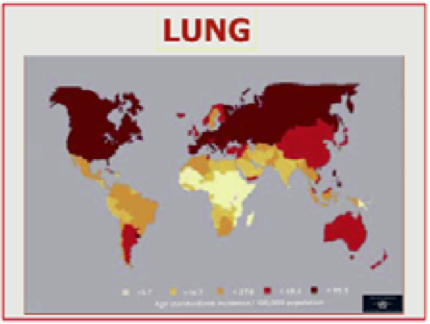


Figure 10.

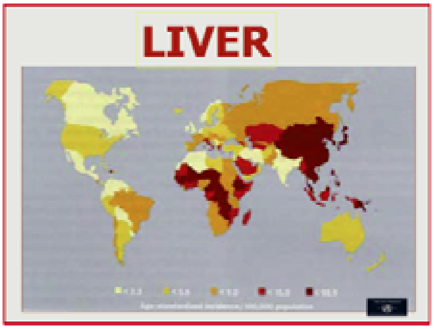


Figure 11.

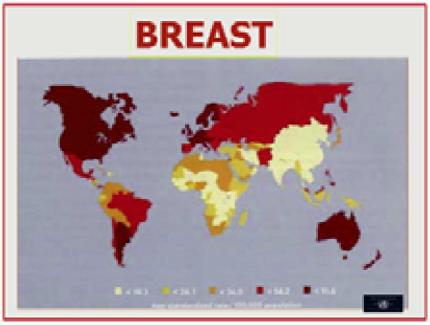
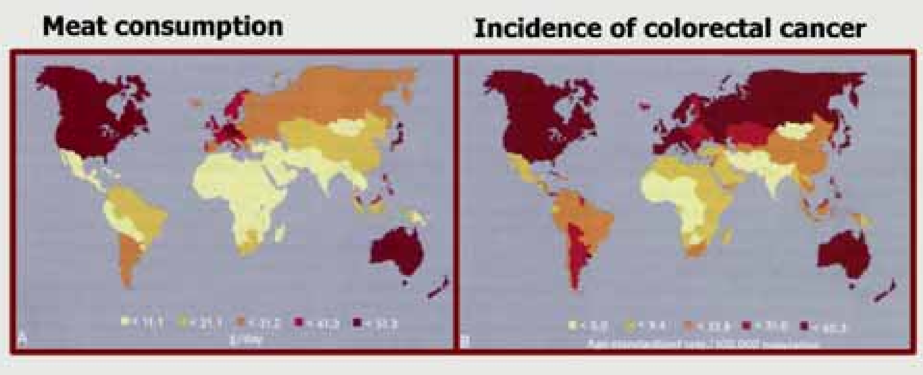


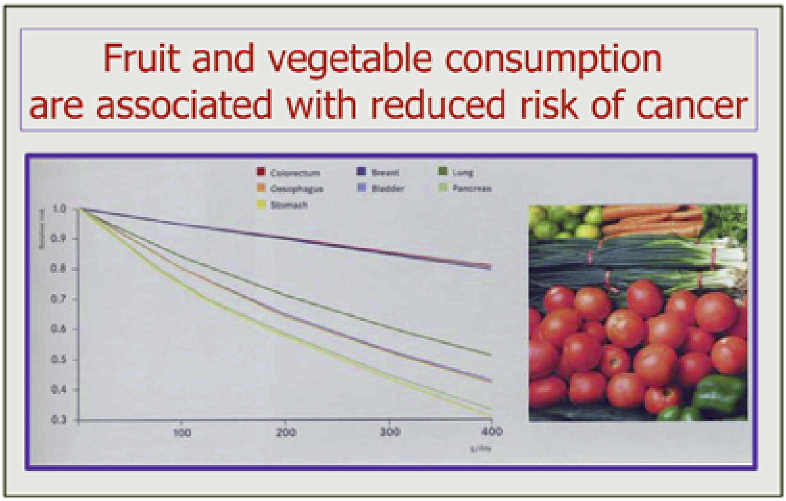
Figure 12.



Meat consumption

Incidence of colorectal cancer

Figure 13.



Fruit and vegetable consumption
are associated with reduced risk of cancer

Figure 14.

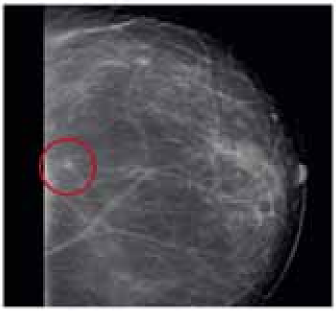


Figure 20.

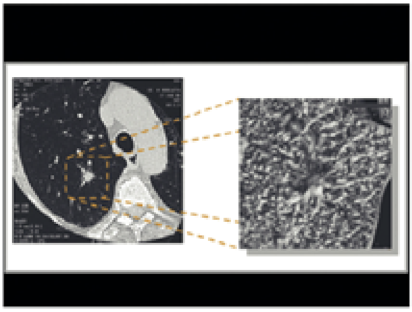


Figure 21.

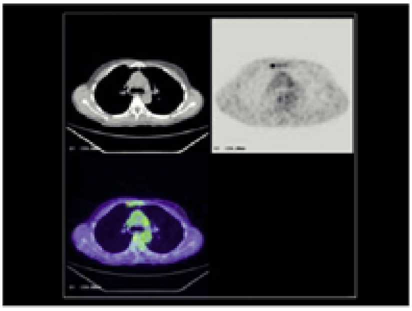


Figure 22.

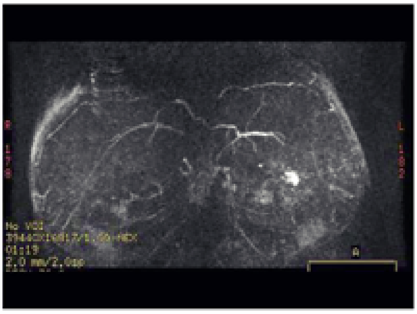


Figure 23.

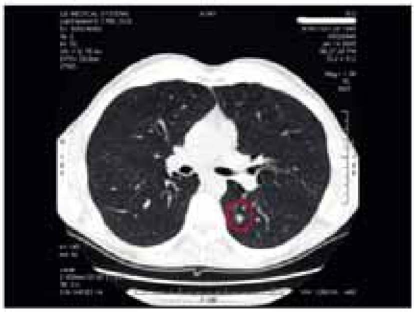


Figure 24.