5

NUMBERS RUNNING WILD

Jeroen P. van der Sluijs

Introduction

This chapter[1] is about craft skills with numbers and, in particular, about problems with the use of numbers of unknown pedigree. As an example, I will discuss a very striking number that appeared in the mainstream press in early May 2015: a new scientific study was reported to have found that 7.9% of species would become extinct as a result of climate change.[2] What was quite remarkable about this number is that it had two digits: not 10%, not 8%, but precisely *7.9%* of species were to suffer for humanity's carbon sins. A question we will address later is

[1] This chapter is adapted from a talk given at "Significant Digits: Responsible Use of Quantitative Information", a workshop organized by the Joint Research Centre of the European Commission at the *Fondation Universitaire* in Brussels on 9-10 June 2015. See https://ec.europa.eu/jrc/en/event/conference/use-quantitative-information

[2] For instance, in the *New York Times*: "Overall, he found that 7.9 percent of species were predicted to become extinct from climate change" (Zimmer, 2015).

the kind of distribution we might expect around this figure of 7.9%.

When a number from the domain of science is catapulted into a user community (for instance, a community of policymakers or conservation campaigners) that has no idea where it came from and no idea of its pedigree, a great deal can go wrong. We have a classic example of this from climate change research, with the concept of climate sensitivity (Van der Sluijs *et al*., 1998).

Climate sensitivity is a metric for what happens to the Earth's average surface temperature if the atmospheric CO_2 concentration doubles compared to pre-industrial times. There were some model calculations in the early 1980s, with one model giving an outcome of a 1.5°C rise in temperature, another giving 4.5°C, and several models giving intermediate outcomes. This range of outcomes was passed over to the impact assessment community (who do calculations on, for example, rises in sea level), who had no idea where the range came from and therefore failed to fully grasp its meaning. They interpreted the range as a confidence interval—specifically, as a 95% interval with a best and a worst case—so the challenge was thus to protect human societies against a rise in sea level that corresponded to a worst-case temperature rise of 4.5°C.

However, that range had nothing to do with a confidence interval; it was simply a range of point outcomes of individual climate models whose reliability was completely unknown. We will come back to this later. However, imagine what would happen if the 7.9% species extinction rate was presented confidently at the next United Nations climate change summit. What should negotiators and policymakers do with such a number?

Perhaps the negotiators would agree to set an 'acceptable' rate of extinction. A typical political compromise might be half of what would be expected in a business-as-usual scenario. The quantitative policy target would thus be set at a maximum 3.95% (half of 7.9%) extinction rate. The level of precision has now moreover increased to an inconceivable three significant digits, which, given the gross absence of numerical craft skills and literacy, is not uncommon at the science–policy interface.

The rest of the present chapter is structured in three parts. The first concerns the background to the enterprise of producing numbers for policy and the problems and challenges involved. Then we will briefly introduce the NUSAP method, which is a systematic way of exploring the unquantifiable parts of uncertainty associated with these types of numbers. The final section of the paper will be about the case of extinction risk from climate change, to go into a little detail about the kind of number it is and where it comes from.

The phenomenon of scientific uncertainty

We commonly find ourselves in real-world situations that can be called 'post-normal'; that is, where decisions are urgent, stakes are high and values are in dispute (Funtowicz and Ravetz, 1993). In such situations we cannot afford to wait—we are compelled to make immediate decisions based on very imperfect information, marked by irreducible and largely unquantifiable uncertainty. The values of the various stakeholders may be in conflict. Usually, the way we produce knowledge in situations like this is quite different to how it is done in the mono-disciplinary natural sciences. There is no reproducible laboratory experiment or measurement in the field, so we typically use simulation models with future scenarios, based on all kinds of assumptions and with very serious limits to validation and even to parameter estimation. There are often many hidden problems in these models that have not yet been systematically exposed or systematically critically reviewed. We thus need more of what we call 'knowledge quality assessment methods'.

Let us discuss two ways of depicting uncertainty in the climate sciences. The first is to imagine a cascade of uncertainties in the causal chain of climate change. It might start with the drivers, such as population growth, *etc.*, and then progress to energy futures (what kinds of energy will be used to meet the energy demands of the future generations), leading to all kinds of fuel mixes, each with different levels of CO_2 emissions. There are also the non-CO_2 greenhouse gases whose sources change over time, which then lead to changes in emissions of these greenhouse gases, while their atmospheric fate depends on complex atmospheric chemistry in which the presence of one greenhouse gas can impact on the atmospheric half-life of another. A part of the CO_2 and non-CO_2 greenhouse gas emissions remains in the atmosphere and the rest is redistributed, so we need some modelling of carbon cycles and atmospheric chemistry, adding another layer of uncertainty to the cascade. That then produces some radiative forcing of the climate system; and there is also climate sensitivity, which is still to a large degree uncertain. We can then do some impact modelling. We also often need regional projections to inform regional and local decisions on adaptation, with all the challenges and uncertainties associated with down-scaling—and so on.

Moving in this way, from population growth to energy futures to carbon emissions to climate change to impacts such as the loss of biodiversity, seems to imply that we know the structure of the models in each step of the cascade—that it is just a matter of inputting the right parameter values in order to do the right calculations and even to quantify the uncertainty in the model-chain outcome for local climate impacts. However, do we really know the structure of this complex system well enough to make a reliable conceptual model to predict its behaviour? Are all the assumptions we make to link the models valid?[3]

A second way of understanding uncertainty in earth system modelling presents itself upon reflection on the implications of the geological record of the atmospheric composition of CO_2 and methane over a period of half-a-million years. From the gas bubbles trapped in ice in the Vostok ice core drillings, scientists have been able to reconstruct the composition of the atmosphere in the past and have found that it varied from between 180 to 280 parts per million (ppm) by volume for CO_2. The projections by the International Panel on Climate Change (IPCC) are that this level could rise to 1,100 ppm or so, while we already live in a world in which 400 ppm of CO_2 have been exceeded. That is far outside the range of our present knowledge of this complex system, which means that *the past is no longer the key to the future*. All kinds of new feedback processes may be at work in this system, which has been stressed far beyond its long-term equilibrium. We are basically sailing into *terra incognita*. The late Roger Revelle (Revelle and Suess, 1957) called this "man's greatest geophysical experiment". We simply do not know how the planet will respond under these pressures, yet we continue to produce crisp numbers such as 7.9%.

How does the science–policy interface cope with uncertainties? Two strategies seem dominant: concerned scientists over-sell certainty to support political action, while other groups in society over-emphasise uncertainty to delay or prevent it. Scientists are afraid that if they are too open about uncertainties, there will be a policy stalemate, and they will furthermore be reprimanded for not having done their homework correctly. Policymakers generally

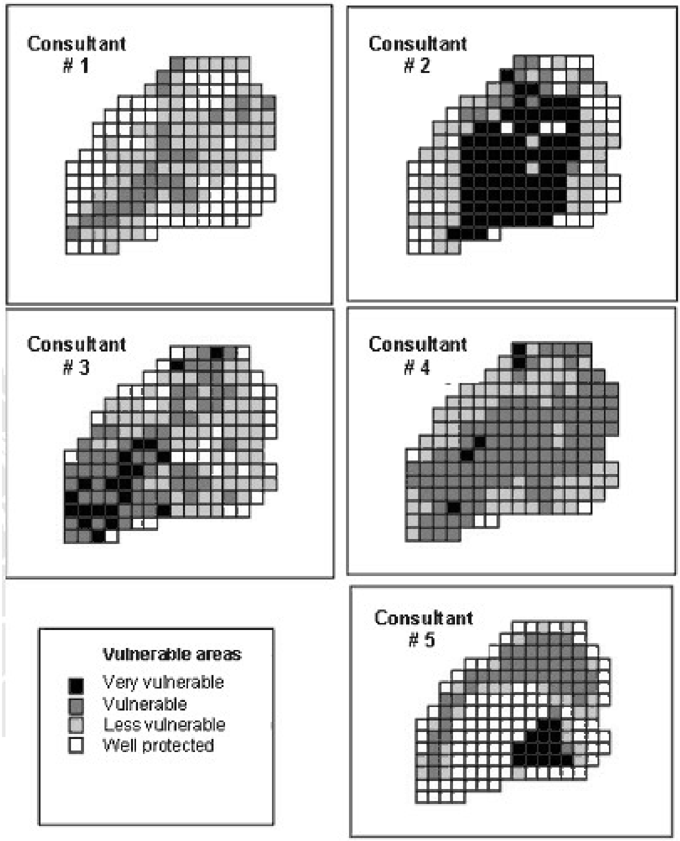
[3] On this point, see, for example, Van der Sluijs and Wardekker (2015) and Saltelli *et al.* (2015).

expect that, with enough hard work, science can and will produce the ultimate answer.

This is the common *paralysis by analysis* pitfall. To avoid it, scientists have started to be less open about uncertainty and to exaggerate certainty and consensus in order to advance political decision making. Both overselling certainty and exaggerating uncertainty are strategies driven by a compulsion to influence the policy process. The problem is that neither strategy is adequate to the challenges posed by the nature of the uncertainties we face. As research progresses, it produces surprises: the more we know, the more we are aware of complexities in the system, which may completely change what we expect in the future. We may therefore need far more flexible policies than we might imagine on the basis of certainty.

Another famous, almost iconic, illustration of the problem of uncertainty is found in a real case concerning the protection of a strategic groundwater resource in Denmark, near Copenhagen. The five best scientists in the country were given the same question: to determine the most vulnerable part of the area, *i.e.* the part which would need the highest level of protection against nitrate pollution from intensive agriculture and land use practices. All the scientists used the same basic data and sets of measurements from the area, but they each used different models and approaches for their assessment.

Figure 1. Model predictions on aquifer vulnerability towards nitrate pollution for a 175 km² area west of Copenhagen



Source: Copenhagen County (2000); Refsgaard *et al.* (2006).

The five vulnerability maps in Figure 1 show the results. Unfortunately, they are not in agreement; they are, in fact, contradictory. However, imagine that science has spoken and that it is now up to policymakers to make a wise decision. What can they do? They can say, "Let us be precautionary and assume the worst case, since if we guard against the worst case, then we are more or less sure that we can protect the zone". But the cost of that strategy may be disproportionately high. They may say, "We need more information before we can decide". That is the infamous *paralysis by analysis* pitfall: more information produces more contradictions, so the decisions are postponed in an infinite loop of evidence-gathering and indeterminate analysis.

So the best science can produce a plurality of scientific perspectives, based on various scientific models and styles of reasoning. It seems that we need to understand better what uncertainty is and why outcomes do not converge. If we look carefully at what is going on in the science–policy interface, we see that the phenomenon of uncertainty is understood in three different ways (Van der Sluijs, 2012). The first is the deficit view, in which uncertainty is seen as a temporary problem. For the time being the science is imperfect, and all we need to do is to collect better data, improve our models, and then, ultimately, science will provide certainty. That is, we reduce uncertainties until we have the precise answer. If that does not work, we just quantify the remaining uncertainty by some confidence interval, error bar, or similar. We use statistical tools such as Monte Carlo simulation, Bayesian belief networks, *etc*. But there is a problem with these tools: we need to feed them with numbers. We need to specify uncertainty ranges for all the model parameters. And if we do not know the range or the distribution and there are no data, we just take a knife, put it to the throat of an expert and ask for a number; we call this 'expert elicitation'. When pushed hard enough, experts will produce a number range or a distribution based on the best of their knowledge, even if such knowledge is illusory. We then put this number range or distribution into the Monte Carlo tool, do the necessary calculations and produce a quantified outcome which fits nicely in a spreadsheet. This corresponds to the 'speaking truth to power' model of interfacing science and decision making. It assumes that we need to produce a quantitative answer, because that is what we believe science is able and supposed to do. And where there is uncertainty, we just speak truth-with-error-bars to policy.

The second understanding of uncertainty is the evidence evaluation view. This is a more pragmatic approach to the problem that science speaks with multiple voices. We see this, for instance, in the way the IPCC works. Experts from different disciplines are brought together to work on a consensus, which means finding the highest common denominator in the science, the most robust claims that we can support with all the published and peer-reviewed studies to date. That is, we know we cannot achieve truth, so we substitute it with a proxy for truth, which is scientific consensus. The model of interfacing science and decision making here is 'speaking consensus to power and to policy'.

These two views can work and be successful in some problem domains, but there are classes of problem where they manifest serious limitations. The first view has the limitation that not everything can be quantified and that experts are forced to quantify uncertainty through best-guessing, because there is no other way. There are many other deficiencies due to model structure uncertainty, the simplification of complex reality in simulation models, the impact of the choice of system boundaries on the outcome, which can never be captured in numeric ranges, and even the assumptions modellers inescapably have to make (*e.g.* Van der Sluijs *et al.*, 2008). The second view has the problem that the consensus approach ignores the unknown probability of high impact scenarios in the risk assessment and tends to confuse unknown probability with negligible probability (*e.g.* Van der Sluijs, 2012). In the first IPCC report, for example, the policymakers' summary mentioned neither the possible collapse of the Antarctic ice sheet nor the collapse of thermohaline circulation in the oceans as relevant scenarios for climate policy. This was not because there was no knowledge; indeed, several published studies had identified those non-linear risks. The detailed chapters of the first IPCC report did review this literature, but it was absolutely impossible to achieve any robust conclusions on these scenarios, so they were excluded from the consensus and did not make it into the policymakers' summary. However, such scenarios were and still are extremely policy-relevant, because policymakers should be interested in the possibility that a five-metre sea level rise could result if the West Antarctic ice sheet were to collapse. Policymakers and society need to know about these types of scenarios even if scientists cannot reach consensus on them.

We therefore need another model, which is the post-normal or complex system approach to uncertainty, in which we acknowledge that uncertainty is permanent and intrinsic to complex, open-ended systems, and that uncertainty can in fact be the result of the way we produce knowledge. For instance, when we construct and use computer simulation models, we cannot avoid making assumptions. There is no way of getting reality into computer code other than by making assumptions, the validity of which can never be thoroughly checked. We thus need another means of controlling the quality of assumptions in models. That requires a more open way of dealing with the deeper dimensions of uncertainty, *etc.*, and is why we have developed tools for knowledge quality assessment (Van der Sluijs *et al.*, 2008, Saltelli *et al.*, 2013).

From this post-normal perspective, the whole science–policy interface changes. It is no longer speaking truth or consensus to power; it is now working deliberatively within imperfections. We cannot always produce the ultimate answer; rather, we have to work within imperfections and we have to do that in dialogue between science, policy and society.

By studying the science–policy interface in different fields—*e.g.* air quality, climate change, biodiversity—we see how differently these types of problems may be solved. A Bayesian approach is often used, where the five different results in the Danish groundwater resource case described above are seen as informative priors that are simply averaged and the resulting model updated if new information becomes available. But what if there is no data, and decisions are needed urgently? What is the quality of the priors? We could take the IPCC approach, in which scientists are locked in a room and not released until they reach a consensus. We could take a precautionary approach and assume the worst case. We could take an academic/bureaucratic approach, whereby the scientist with the highest Hirsch index is given the greatest credibility. We could give preference to the scientist we trust most or the one most in line with our policy agenda, which is what often happens. We could forget the science and decide on an entirely different basis. Alternatively, we could explore the relevance of our ignorance in a post-normal attitude, whereby we try collectively to find wiser ways to deal with uncertainties and avoid the pitfall of taking decisions under the illusion of having tamed uncertainty.

The first view of uncertainty as a temporary state has been very persistent. In a quotation from the first IPCC report in 1990 (IPCC, 1990; see Box 1a), the authors write that there are many uncertainties in their predictions, particularly with regard to the timing, magnitude and regional patterns of climate change due to incomplete understanding of sources and sinks of greenhouse gases, clouds, oceans and polar ice. They go on to say that these processes are already partly understood, and that they are confident that the uncertainties can be reduced by further research. That was a strong claim in 1990 for the scientific community working on climate change; it is what they believed.

Box 1.

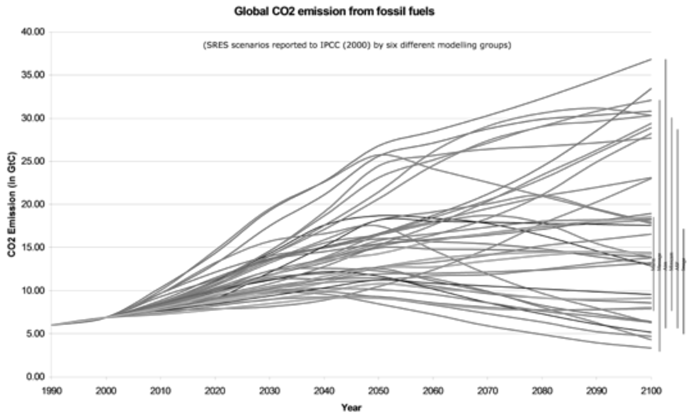
Box 1a. IPCC 1990 optimism about reducing uncertainty.	**Box 1b. Former IPCC chairman, the late Bert Bolin, on the objective to reduce climate uncertainties.**
"There are many uncertainties in our predictions particularly with regard to the timing, magnitude and regional patterns of climate change, due to our incomplete understanding of: • sources and sinks of greenhouse gases, which affect predictions of future concentrations • clouds, which strongly influence the magnitude of climate change • oceans, which influence the timing and patterns of climate change • polar ice sheets which affect predictions of sea level rise These processes are already partially understood, and we are confident that the uncertainties can be reduced by further research However, the complexity of the system means that we cannot rule out surprises." (IPCC, 1990: xii)	"We cannot be certain that this can be achieved easily and we do know it will take time. Since a fundamentally chaotic climate system is predictable only to a certain degree, our research achievements will always remain uncertain. Exploring the significance and characteristics of this uncertainty is a fundamental challenge to the scientific community." (Bolin, 1994)

Just before the second IPCC assessment report (1995) was issued, the late Bert Bolin, then chairman of the IPCC, said in a conference that they could not be certain that uncertainty could be easily reduced. He continued by saying it would take time to reduce uncertainty, in line with the deficit view of temporary nature of uncertainty. But he concluded that, since the fundamentally chaotic climate system is predictable only to a certain degree, there were limits to predictability, and research results would always remain uncertain. He thus finished by adopting the third view we have identified, namely that uncertainty is permanent and will never go away.

Bolin added that exploring the significance and characteristics of this uncertainty was a fundamental challenge to the scientific community. Most people would say that the fundamental challenge to science is to determine facts and achieve certainty, so in this respect Bolin was being 'post-normal', even in the early stages of post-normal science.

Let us look more closely at the IPCC's ambition in 1990 to reduce uncertainty in various domains (Box 1a). The first item they listed was "sources and sinks of greenhouse gases". Ten years later, what we know about sources and sinks of CO_2, only one of various greenhouse gases, is illustrated in Figure 2.

Figure 2. Emission scenarios produced by six different energy models (Maria, Message, Aim, Minicam, ASF, Image), each using four harmonized sets of scenario assumptions, as presented in the IPCC Special Report on Emissions Scenarios



Source: IPCC (2000).

Four different sets of assumptions, A1, A2, B1 and B2, were used in each of six models presented in the IPCC Special Report on Emissions Scenarios (2000), corresponding to a focus on economy (A) versus environment (B) and on development and globalization (1) versus regionalization (2) of the world. These scenario families ensured harmonized assumptions on population growth, economic growth, and technology transfer. Despite this, there were very wide discrepancies in what the six models projected for these same four sets of scenario assumptions. Apparently not much has been reduced in terms of uncertainty about sources and sinks of CO_2 in ten years of research.

Now let us look at the key parameter of climate change, so-called climate sensitivity, 25 years after the first IPCC report predicted a reduction in uncertainties. As indicated above, climate sensitivity is a measure of the average global warming that would be produced by a doubling of the CO_2 concentration (to 560 ppm) compared to the pre-industrial level of 280 ppm. The first synthesis estimate of climate sensitivity was made in 1979, when the U.S. National Academy of Sciences made an assessment that combined results from several general circulation models. The outcomes ranged from 2 to 3.5°C; with some expert wisdom the range was widened, in acknowledgement that the models were imperfect. By including this tacit knowledge of uncertainties in the models, they arrived at the conclusion that the anticipated average rise in temperature must be somewhere between 1.5 and 4.5°C, and their best guess was 3°C.

Table 1. Evolution of knowledge on climate sensitivity over the past 35 years

Assessment report	Range of GCM results (°C)	Concluded range (°C)	Concluded best guess (°C)
NAS 1979	2-3.5	1.5-4.5	3
NAS 1983	2-3.5	1.5-4.5	3
Villach 1985	1.5-5.5	1.5-4.5	3
IPCC AR1 1990	1.9-5.2	1.5-4.5	2.5
IPCC AR2 1995	MME	1.5-4.5	2.5
IPCC AR3 2001	MME	1.5-4.5	Not given
IPCC AR4 2007	MME	2.5-4.5	3
IPCC AR5 2013	MME (0.5-9)	1.5-4.5*	Not given

Source: Updated from Van der Sluijs *et al.* (1998).

* “Likely” (17-83%) range. Note that prior to Assessment Report 4 (AR4), ranges were not clearly defined. MME = Multi Modal Ensemble.

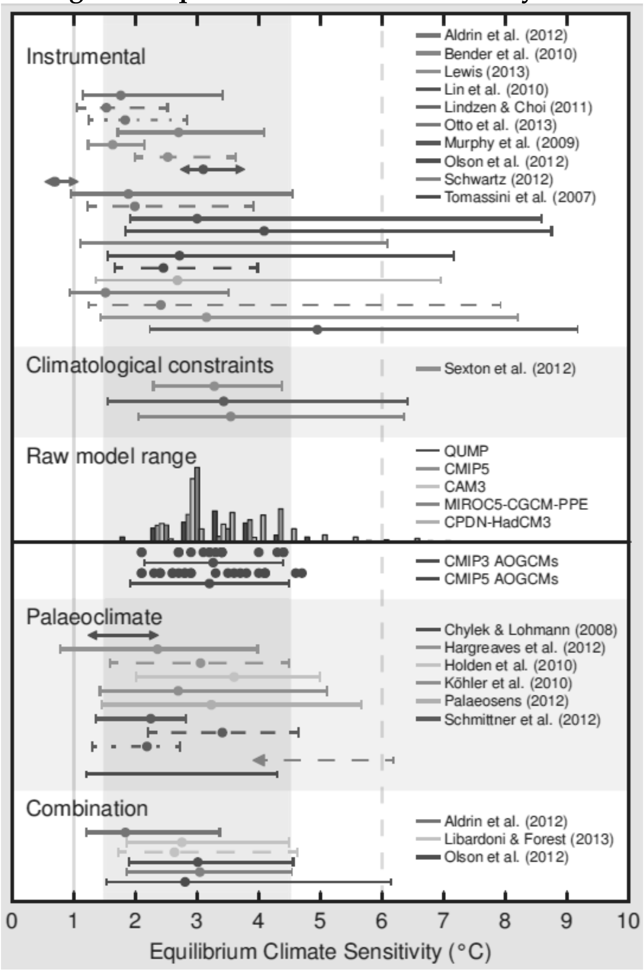
A couple of years later the National Academy of Sciences (NAS) updated the assessment with new outcomes from models; thereafter came the Villach conference in 1985, where the published literature showed models with outcomes as low as 1.5°C and as high as 5.5°C (see Table 1). However, something strange then occurred, because the range given in the conclusions of the Villach conference was narrower than this. The NAS had widened the range in their assessment because the experts took into consideration the uncertainties in the model. But in the new assessment in Villach in 1985, the range resulting from the inventory was narrowed by the experts, because they did not want to grant too much credibility to one (perceived) outlier model. In effect, the argumentative chain linking the set of published literature to the recommended range completely absorbed all the changes in the science and all the uncertainties, to keep the scientific basis for policy making stable.

From the second IPCC report onwards, multi-model perturbed physics ensemble modelling was used, in which each model produced a range or distribution which was then combined into an ensemble. The fifth report (IPCC, 2013) suddenly has a footnote on the recommended range, so the multi-model ensemble produces numbers between half a degree and 9°C for climate sensitivity; the recommended range is still 1.5 to 4.5°C; and the footnote indicates that this is to be interpreted as the likely range, the 17^{th} to 83^{rd} percentile of the distribution.

The modellers involved in the first assessment report said in an interview that the range recommended in that report was to be seen as the equivalent of a 95% interval (Van der Sluijs *et al.*, 1998), whereas in the fifth report that same 1.5-4.5°C range has suddenly become a 66% interval. In other words, the meaning of the range has changed in order to keep it constant over time for use in the science–policy interface.

Looking at the distributions from different sources (models, paleo records, *etc.*) (Figure 3), we see some ranges, such as the likely 1.5 to 4.5°C range from Chapter 12 of the fifth assessment report of the IPCC. More than a century ago, Svante Arrhenius (1896) did the first calculations on the potential doubling of CO_2 and came up with 4.95°C at the Equator, rising to 5.95°C at 60°S and to 6.05°C at 70°N. The IPCC's fifth assessment says the temperature rise is very unlikely to be greater than 6°C, so even the one-hundred-year-old estimate without computer models, with crowd sourcing to students for the calculations, is still well in the range of what IPCC now considers possible. Have uncertainties diminished, as the IPCC foresaw in 1990? On the contrary, they seem to have increased, at least in this key climate parameter.

Figure 3. Probability density functions, distributions and ranges for equilibrium climate sensitivity

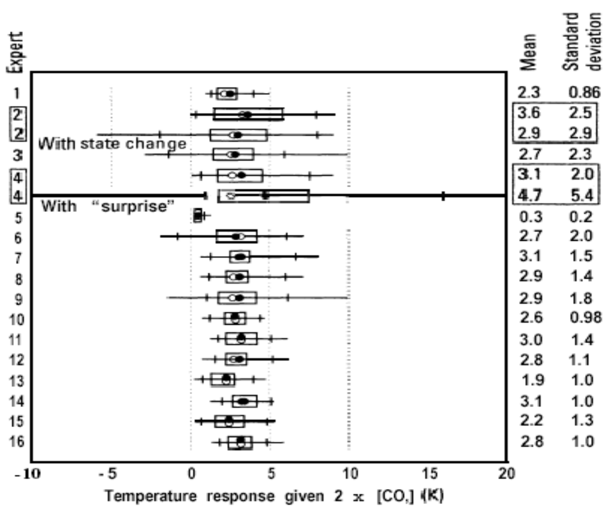


Grey shaded range: likely 1.5°C to 4.5°C range. Grey solid line: extremely unlikely, less than 1°C. Grey dashed line: very unlikely, greater than 6°C.

Source: Reprinted from Box 12.2, Fig. 1, from Collins *et al.* (2013).

In addition to asking the models, one can also ask the experts, as Morgan and Keith (1995) did. Looking at subjective assessments of climate sensitivity (Figure 4) by the top 16 climate modellers (based on their publication record in the field) in the USA, we see that some cannot even exclude the possibility that it is a negative number—that is, that the equilibrium temperature change will be smaller than zero due to some feedback that overcompensates for the warming by cooling in the long run. Overall, we see that there are many ways of assessing climate sensitivity that lead to very different views.

Figure 4. Box plots of elicited probability distributions of climate sensitivity, the changes in globally averaged surface temperature for a 2 X [CO_2] forcing



Horizontal line denotes range from minimum to maximum assessed possible values. Vertical tick marks indicate locations of lower 5th and upper 95th percentiles. Box indicates interval spanned by 50% confidence interval. Solid dot is the mean and open dot is the median. The two columns of numbers on the right side of the figure report values of mean and standard deviation of the distributions.

Source: Morgan and Keith (1995).

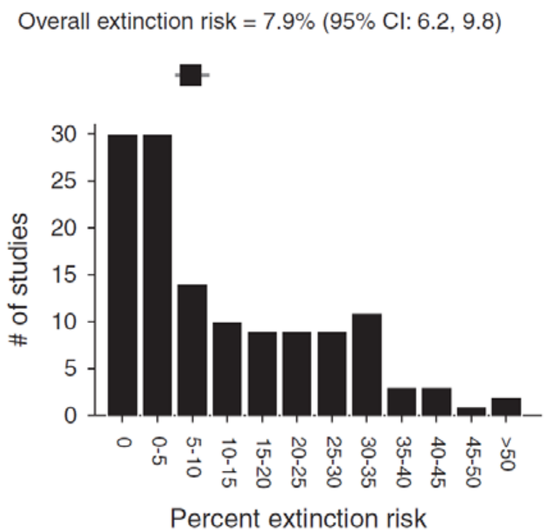
The case of species extinction from climate change

We return now to the number 7.9% with which this paper began, namely the predicted rate of extinction of species due to climate change. Models have produced widely varying estimates of the anticipated extinction of species. Urban (2015) assembled 131 studies. At first sight, 131 is an easily understandable number. It is just a count of the number of studies. But even that number is problematic, because some studies combined a number of other studies, but were counted as one, while there was also some overlap due to the fact that any given study may have combined several studies that were considered separately elsewhere, and so on.

Urban's meta-analysis, published in *Science*, found that, overall, these 131 studies predicted that 7.9% of species would become extinct due to climate change. He also provided a confidence interval: for 2°C it would be 5.2% and for 4.3°C it would be 16% of all species. Linguistically Urban is literally talking about a percentage of "all species". The opening paragraph of the study that presented this 7.9% tells us that the goal of the whole exercise was to inform international policy decisions about the biological cost of failing to curb climate change and to support specific conservation strategies to protect the most threatened species.

In practice, here is what Urban did. He considered 131 studies and classified them according to their outcome in terms of predicted risk of extinction, with one category of zero (containing 30 studies), another for zero to 5% (containing another 30 studies), then 5% to 10% (containing about 14 studies), *etc*. Then we see an average, 7.9%, summarizing all of these studies. It is not clear what year this number refers to. Each of the 131 studies had a different reference period, so no year is mentioned in the paper. Zero to 5% includes zero, so one should stack the two bars for the first category in Figure 5 (60 studies), making for a very asymmetrical distribution. The figure 7.9% does not give any useful information about this distribution at all. It is misleading to communicate the average of such a skewed distribution without also giving the distribution; in addition, we should know where all these numbers come from.

Figure 5. Histogram of percent extinction risk from climate change for 131 studies



Source: Urban (2015).

Systematic critical reflection on uncertainty

Before moving on to another estimate of climate extinction risk from Thomas *et al.* (2004) in *Nature*, let us look at how we should critically reflect on uncertainties in these types of studies. This is the guidance approach (Figure 6) that the Netherlands Environmental Assessment Agency developed jointly with a trans-disciplinary group of scientists in the field of uncertainty, with backgrounds ranging from the policy sciences and the humanities to the natural sciences. It starts from the idea that we need critical reflection on all the phases of knowledge production, from the framing of the problem and the drawing of the system boundaries in making our models, to what is included and excluded—whether, for example, we should treat climate change separately from ocean pollution, air pollution, *etc.*, or rather try to include all these complex interactions.

Figure 6. The Dutch Guidance approach to systematic reflection on uncertainty and quality in science for policy

Foci	Key issues
Problem framing	Other problem views; interwovenness with other problems; system boundaries; role of results in policy process; relation to previous assessments
Involvement of stakeholders	Identifying stakeholders; their views and roles; controversies; mode of involvement
Selection of indicators	Adequate backing for selection; alternative indicators; support for selection in science, society, and politics
Appraisal of knowledge base	Quality required; bottlenecks in available knowledge and methods; impact of bottlenecks on quality of results
Mapping and assessing relevant uncertainties	Identification and prioritisation of key uncertainties; choice of methods to assess these; assessing robustness of conclusions
Reporting uncertainty information	Context of reporting; robustness and clarity of main messages; policy implications of uncertainty; balanced and consistent representation in progressive disclosure of uncertainty information; traceability and adequate backing

Source: Van der Sluijs *et al.* (2008).

We always simplify in scientific assessments. We tend to set very limited system boundaries in order to keep scientific assessments manageable, but we have to understand the impact of these design choices on the validity and scope of the conclusions of such assessments. Stakeholder engagement is ever more important. Stakeholders are crucial in co-framing the problem and co-deciding what is relevant to address; they can provide useful information and data that scientists have otherwise no access to; and they can be a critical resource in quality control and extended peer review.

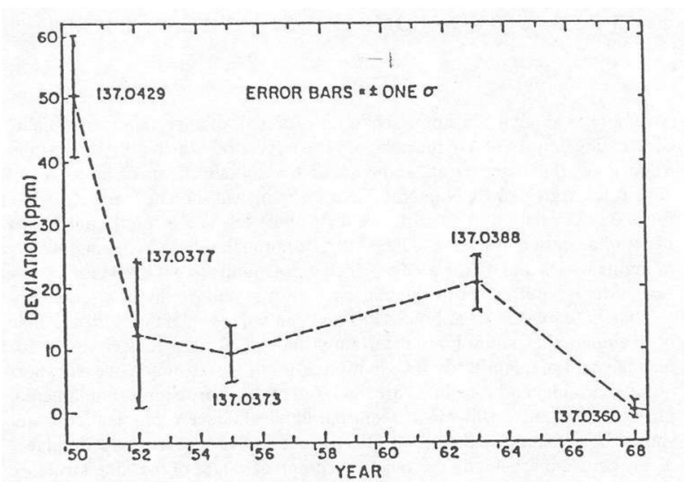
Next, there is the selection of indicators. In scientific assessments all kinds of indicators are used: for instance, the percentage of species that are at risk of extinction or the anticipated temperature rise due to the doubling of CO_2 concentrations. Are these indicators relevant to the policy challenges that we face? Are there alternatives? Are they chosen because a particular model is available to use, or were they designed to address a particular societal problem? Usually, scientists attempt to answer the questions of policymakers by using an existing model or toolkit that is on the shelf but which does not really match the decision-making needs.

Then there is the appraisal of the knowledge base, meaning that we have to systematically look at all the problems and limitations in the available knowledge and characterize it in terms of its pedigree, strengths and weaknesses. We will return to the concept of pedigree later.

Next, we have to map all the sources and relevant types of uncertainty, to be aware of where they are, what they are, and how to take them into consideration in the policy advice; we then have to report the information on uncertainty in a way appropriate to the decision-making context, so that decision-makers can actually use it to make more robust, more resilient or more flexible decisions that take these uncertainties into account in a more sophisticated way than is currently done.

The core message is that uncertainty is much more than a number and that it has dimensions. Three of these are defined in the book *Uncertainty and Quality in Science for Policy,* the 1990 classic by Funtowicz and Ravetz. There are technical, methodological and epistemological dimensions of uncertainty in numbers. The technical dimension corresponds to inexactness, error bars, *etc.*, while the methodological dimension corresponds to unreliability (for instance, whether the model structure is reliable). The epistemological dimension corresponds to the borders with ignorance and the limits of our ability to know and understand complex systems. We could argue that there is another dimension, namely the societal robustness of the knowledge, but this could also be considered to belong to the epistemological category.

Figure 7. Successive recommended values of the fine structure constant α^{-1}



Source: Taylor *et al.* (1969).

An example from Funtowicz and Ravetz (1990) concerns the recommended number for a physical constant, the *fine structure constant*. Recommended values reported in successive editions of the *Handbook of Chemistry and Physics*—the bible of the natural sciences—have changed over time. This is because science progresses. There are three items in the notational system used in the *Handbook*:

a number, a unit and a spread, the latter usually reported as a standard deviation. If we look at the number and the standard deviations that are recommended in successive editions of the *Handbook* for the fine structure constant, we see something strange.

If there were only measurement error with random distribution to be considered, we would expect that 95% of the distribution would be captured by two standard deviations around the mean, and almost everything captured by plus or minus three standard deviations. However, we see that in 1968 the recommended value and reported error bar (one standard deviation) are far outside the error bar reported in 1950; the recommended values in 1950 are more than four standard deviations apart. So apparently there is more to say about uncertainty than can be captured in this standard deviation.

This was perhaps the first time the fine structure constant had been measured, but we know that systematic error means that no two laboratories will produce exactly the same result. If users of this number could have been warned early on that it had some limitations and that the reported standard deviation could not capture all the uncertainty, it might have helped to avoid misunderstandings of this type of number. This kind of uncertainty has somehow to be communicated.

This example was of a physical constant, which may not capture the imagination of the public. Let us consider a policy relevant number, the emission levels of ammonia, an air pollutant mainly coming from the agricultural system and for which emission reduction policies are in place in the Netherlands. Intensive cattle breeding in the Netherlands produces a great deal of NH_3 (ammonia) air pollution. Emission reduction policies are often relative to a reference year. Let us take 1995 as the reference year, and imagine that we want to reduce emissions by, for instance, 10% relative to that year. We take the Dutch "State of the

Environment Report"[4] (RIVM, 1996, 1997, 1998, 1999, 2000, 2001, 2002), which has tables for all the environmental indicators for which the Netherlands has implemented policies. We look up the 1995 emissions levels in the 1996 edition of the report, and it is slightly more than 150 million kilograms. Then we look up the same number in more recent editions of the report. In the 1997 edition there were some recalculations and there is a new number for the 1995 emission, which is slightly lower (see Figure 8). Nothing changes in the 1998 edition, but there is a jump in the 1999 edition: it seems that knowledge of the emissions in 1995 has suddenly changed, and we see now that the 1995 level is almost 30% higher than we thought a couple of years earlier. The 30% change in the number due to the recalculation is three times larger than the hypothetical long-term policy objective of 10% reduction in emissions, which is discomfiting because it has serious policy implications.

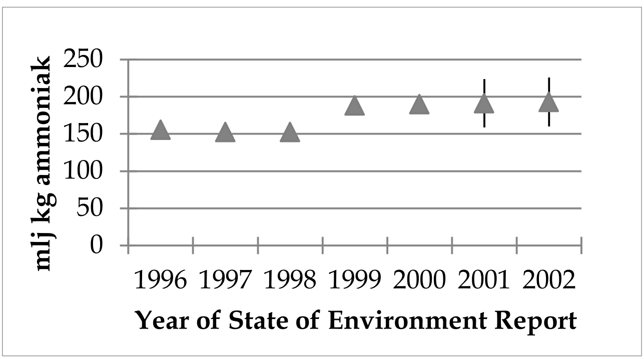
What happened here? We might imagine that someone goes out to measure gas emissions every year, but that is not the case. The calculation of NH_3 emissions in a given year is the product of an agricultural model that follows the nitrogen, starting with the animal feed and taking into account the number of cows, chickens and pigs in the Netherlands. These animals are kept in different agricultural subsystems, some in barns, some in meadows. The model assumes partition coefficients of how much nitrogen from the food ends up in the manure. A fraction of the nitrogen in the manure in a barn or in a meadow is emitted as NH_3 into the air. This is then calculated for each type of cow, chicken and pig in each agricultural subsystem and finally everything is summed to serve as an

---

[4] "Milieubalans", an annual publication by the Netherlands Environmental Assessment Agency, http://www.pbl.nl, which was formerly part of RIVM (National Institute for Public Health and the Environment).

estimate of the total NH_3 emissions in a given year. So the numbers of animals in 1995 considered in the calculation remained constant, and what changed in the recalculation in the 1999 edition was the assumed emission factor of the biggest type of barns used in the Netherlands. A small change in the emission factor of the largest pig farms makes a big difference in the outcome for the model-calculated emissions.

Figure 8. Total NH_3 emissions in the Netherlands in 1995 as reported in successive editions of the "State of the Environment Report" of the Netherlands Environmental Assessment Agency



Uncertainty as discussed in the present chapter may appear to be the preoccupation of a small community of practitioners. In fact, its relevance to policy making may sometimes cause it to hit the headlines. In 1999 there was a scandal at the Netherlands Environmental Assessment Agency when whistle-blower Hans de Kwaadsteniet revealed to the media the use of poorly validated models for the production of environmental assessments (Van der Sluijs, 2002). As a consequence of the scandal and its follow-up, it became mandatory for the Environmental Assessment Agency to report uncertainties. From the 2001 edition of the "State of the Environment Report" onwards error bars are reported, reflecting two standard deviations. What is striking is that the size of these error bars is smaller than the change observed as a result of the recalculation in the 1999 edition. In terms of the different types of uncertainty, that means that the methodological uncertainty (recalculation with updated assumption) was more significant here than the technical uncertainty (error bar).

Qualified quantities: the NUSAP approach

What about the third dimension, the borders with ignorance, which might be even more important? How do we assess that dimension of uncertainty? We need more qualifiers of scientific information, and that is where the new notational system comes in. The classic notational system only has room for a number, a unit and a spread, whereas for all the problems in the post-normal domain, we may need to issue something like a patient information leaflet that warns users about the risks involved in using numbers and provides guidance on how to use them responsibly. To that end, Funtowicz and Ravetz (1990) proposed adding *assessment* and *pedigree* to the notational system, producing the NUSAP system: numeral, unit, spread, assessment, pedigree. Assessment is an expert judgement on reliability that puts the number into context, and pedigree is a systematic multi-criteria evaluation to characterize the status of the scientific information in terms of how it was produced, where it comes from, and how we should understand the pedigree of a given number. This system is sensitive to our proximity to ignorance. Pedigree is assessed with the help of a pedigree matrix, describing pedigree criteria along with an ordinal 5-point scoring scale for each criterion. Table 2 presents how a pedigree may be articulated.

For example, a model may have different parameters, some based on well-established science and others more speculative, and we have to be able to make the distinction between them. In terms of criteria we can use, we have the empirical basis for the number, the theoretical basis behind the parameter, the methodological rigour with which the number was produced, the degree of validation or the degree to which it was a proxy as compared to the thing we really want to know.

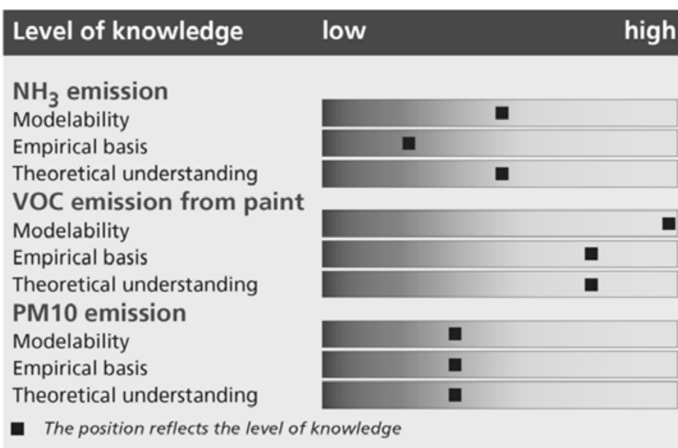
Table 2. Example pedigree matrix for emission monitoring data

Score	Proxy	Empirical basis	Methodological rigour	Validation
4	An exact measure of the desired quantity	Controlled experiments and large sample, direct measurements	Best available practice in well-established discipline	Compared with independent measurements of the same variable over long duration
3	Good fit or measure	Historical/field data uncontrolled experiments, small sample, direct measurements	Reliable method common within est. discipline, best available practice in immature discipline	Compared with independent measurements of closely related variable over shorter period
2	Well correlated but not measuring the same thing	Modelled/derived data, indirect measurements	Acceptable method but limited consensus on reliability	Measurements not independent, proxy variable, limited period
1	Weak correlation but commonalities in measure	Educated guesses, indirect, approx. rule of thumb estimate	Preliminary methods, unknown reliability	Weak and very indirect validation
0	Not correlated and not clearly related	Crude speculation	No discernible rigour	No validation performed

Source: Van der Sluijs *et al.* (2005); adapted from Ellis *et al.* (2000).

For example, the empirical basis can be anything between crude speculation, an educated guess, model-derived data, small samples of direct measurements or, in the best case, a large sample of direct measurements of the thing we seek to know. The level of methodological rigour can range from a state of no discernible rigour to a preliminary method with unknown reliability, all the way to the best available practice in a well-established discipline. Validation can go from weak or indirect validation all the way to comparison with independent measurements. If we have this type of meta-information about the numbers, we can understand that they are not as strong as they might look at first sight and are much more prepared for them to change in the future.

Figure 9. Example of presentation of pedigree scores of monitoring data of air pollutants.



Source: Van der Sluijs *et al.* (2008). © IOP Publishing Ltd. CC BY-NC-SA, doi: 10.1088/1748-9326/3/2/024008

To illustrate, three air pollutants are monitored in a country, but the status of knowledge on how we can best estimate the yearly emissions varies. We have much more knowledge about the volatile organic components of paint, which we can easily model and for which the empirical basis and theoretical understanding are quite strong, than about how particles are formed in a combustion process or about the empirical basis for monitoring ammonia emissions from cattle breeding. This type of information gives a better idea of what types of policies can be formed on the basis of these numbers. Figure 9 gives an example of how this can be communicated.

A paper on extinction risks due to climate change from Thomas *et al.* (2004) serves as another example. It is a four-page paper in *Nature* presenting a very simple model. The study's authors predict, on the basis of a mid-range climate warming scenario for 2015, that 15% to 37% of species in the sample of regions and taxa will be "committed to extinction". This was one of the studies included in the meta-analysis that arrived at the 7.9% figure. The 7.9% in Urban (2015) referred to "species", while Thomas *et al.* refer to "species in their sample of regions and taxa". The meta-analysis talks about species that will become extinct, while this paper talks about species being "committed to extinction", which is a rather vague, poorly defined concept.

There is an interesting claim in the last line of the introductory paragraph of the Thomas *et al*. paper, that these estimates "show the importance of rapid implementation of technologies to decrease greenhouse gas emissions and strategies for carbon sequestration". The logic is not clear. Does the fact that species will go extinct mean that we should design strategies to store CO_2 underground? How do the authors jump from species extinction to a particular preferred solution (carbon sequestration)? This sounds like an opinion for which the underlying arguments are not even given. However, what we are interested in is the quantified outcome of this study, the rate of 15% to 37% of species being committed to extinction under mid-range climate warming scenarios. The authors arrived at this range by applying the so-called species-area relationship from ecology, which says that the number of species depends on the area of the habitat of the species, and that by consequence, if the habitat shrinks, the likelihood of a species becoming extinct increases. The greater the shrinkage of the area, the more species will become extinct. There is a simple relationship between the number of species and the area of a given habitat; its parameters have been estimated based on empirical data:

$$S = c\,A^{z}$$

where S = number of species, A = area, c = constant and $z \approx 0.25$. We see that this formula yields a number, and if we compare it to climate projections, and the area before and after climate change for each habitat, then we get a ratio which is simply the ratio of the area to the power z, which is itself an empirically derived parameter estimated to be 0.25. There was, however, a sensitivity analysis for the value of z.

The authors collected climate projections for habitat changes from the literature and grouped these into three classes of warming (low, mid-range, high) and explored the outcomes for two extreme assumptions on dispersal (no dispersal and full dispersal). To briefly explain dispersal: as a rule of thumb, for each degree of warming we get roughly a 100-kilometre shift of climate zones from the Equator to the poles, a 150-metre upward shift in the mountains. The species have to catch up with that and migrate (disperse) to the new area with favourable climatic conditions for their survival. Some can migrate more easily than others. A forest needs more time to migrate, because trees cannot walk, so it depends on the speed of the seed cycle. From the paleo-record we know that an oak forest can keep up with 0.12°C per decade, and if the increase in temperature is faster, the oak forest will die back more rapidly on one side than it will expand on the other to populate the new zone where the climate is now favourable.

Therefore, the climate zones and associated habitats shift poleward with climate change, but if there is no dispersal, species can only survive in the overlapping area between the old and the new favourable climate zones, while if there is full dispersal, the species can migrate completely to the new, favourable area.

We see that the outcome that made it to the conclusion is the mid-range scenario, with the lowest and the highest number in the reported range corresponding to the full and no dispersal assumptions; the resulting range being 15% to 37%. Next, Thomas *et al*. (2004) present a huge table in which they list in aggregated (per taxa and region) form the data from the 1,103 species which they studied, including mammals in Mexico, birds in different regions, frogs, reptiles, butterflies, plants, *etc*. There are three numbers for each scenario with and without dispersal, representing three ways of aggregating the data from individual species in each taxa and region to the whole set of species in that taxa and region.

However, there are many missing numbers in the table. The authors explain in the methods section how they interpolated the missing numbers, which they had to do in order to calculate the bottom line, the extinction risk for all the species in the sample. These aggregated numbers, which are used in the conclusion of the paper, thus include a large number of hidden interpolated numbers.

Overall, we see that there are many assumptions hidden behind the numbers, for example in the interpolation algorithm. There is also a bias in the sample of species that are included in the analysis, because they are species on which publications exist in terms of what climate change does to their habitat. This means that the study may be focused on those species most sensitive to climate change:

such results cannot be extrapolated to all the species on the planet. Considering again the Urban (2015) study that concluded that 7.9% of all species will be made extinct—what would that be in absolute terms? Nobody knows the number of species on the planet. Does it refer to eukaryotic species only (estimated to be 8.7 million plus or minus 1.3 million (Mora *et al.*, 2011)) or to all species on Earth? Urban does not specify, so even the *unit* (the second qualifier in the NUSAP) is ambiguous.

All these aspects of uncertainty become evident in the process of applying NUSAP and the pedigree matrix to systematize critical reflection on the strengths and weaknesses of scientific knowledge claims. The pedigree matrix can have criteria such as proxy, quality and quantity of the empirical basis of this model, theoretical understanding, representation of the underlying mechanisms, plausibility and colleague consensus. As an exercise, my M.Sc. and Ph.D. students apply such a pedigree matrix to the Thomas *et al.* (2004) study. The estimate of the extinction risk is often scored somewhere in the middle for proxy. There is a correlation between area of habitat size and extinction risk, but it is not the same thing; we are not modelling extinction, but area loss, which is assumed to be correlated to extinction. The quality and quantity of the model, which are somewhere between an educated guess and modelled and derived data, attain quite a low pedigree score of between 1 and 2. Regarding theoretical understanding, it is an accepted theory of a partial nature with a limited consensus on reliability, but some students think it is a preliminary theory. The other pedigree criterion also gets quite low scores, based on preliminary methods and weak and indirect validation.

Conclusions

Figures such as the 7.9% that we have discussed in detail in this chapter are often a first attempt to quantify a complex phenomenon, in this case the risk of extinction of species due to anthropogenic climate change. We have treated this number as an example of what we perceive to be a dangerous practice—the production and use of crisp figures to give the impression that science produces truth.

The fact that such hyper-precise quantifications may emerge from statistical averaging processes or from computational algorithms is not a justification for their publication. Numbers ought to be used responsibly; in our view the approach of Urban (2015) leaves much room for improvement. Science—especially when deployed at the science-policy interface—should involve what we call 'craft skills' with numbers. Quantification should be a much more nuanced and reflective process (Porter, 1995; for a discussion see Chapter 2, this volume). A peer review process for quantitative evidence would need to systematically include approaches such as NUSAP, sensitivity auditing in the case of mathematical or statistical models, and in general the exercise of good judgment.

The liberty to quantify should be used with discretion, and should go hand-in-hand with the duty to refrain from quantification when appropriate. The practice of throwing magic numbers into the arena for public consumption is in our view one of the symptoms and causes of the crisis in science that is under analysis in this book, representing, as it does, an abdication of the scientist's responsibility to ensure that the craft of science has been applied with all due diligence.

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(251): 237-276.

Bolin, B., 1994. "Science and Policy Making". *Ambio* 23(1): 25-29.

Collins, M., Knutti, R., Arblaster, J., Dufresne, J.-L., Fichefet, T., Friedlingstein, P., Gao, X., Gutowski, W.J., Johns, T., Krinner, G., Shongwe, M., Tebaldi, C., Weaver A.J. and Wehner, M., 2013. Long-term Climate Change: Projections, Commitments and Irreversibility. In: Stocker, T.F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S.K., Boschung, J., Nauels, A., Xia, Y., Bex V. and Midgley P.M. (eds.) *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, UK: Cambridge University Press.

Copenhagen County, 2000. "Pilot project on establishment of methodology for zonation of groundwater vulnerability. Proceedings from seminar on groundwater zonation", 7 November. County of Copenhagen (in Danish).

Ellis, E. C., Li, R. G., Yang, L. Z. and Cheng, X., 2000. "Long-term Change in Village-Scale Ecosystems in China Using Landscape and Statistical Methods", *Ecological Applications*, 10: 1057-1073.

Funtowicz, S. O. and Ravetz, J. R., 1990. *Uncertainty and quality in science for policy*. Dordrecht: Kluwer.

Funtowicz, S. O. and Ravetz, J. R., 1993. "Science for the Post-Normal Age", *Futures*, 25(7): 735-755.

IPCC, 1990. "Policymakers' summary", in Houghton, J., Jenkins, G. and Ephraums, J. (eds.), *Climate Change, The IPCC Scientific Assessment*: vii-xxxiv. Cambridge, UK: Cambridge University Press.

IPCC, 2000. *Emissions Scenarios*. Cambridge, UK: Cambridge University Press.

IPCC, 2013. "IPCC Fifth Assessment Report: Working Group I Report", in *Climate Change 2013: The Physical Science Basis*. Geneva: IPCC.

Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G. B. and Worm, B., 2011. "How Many Species Are There on Earth and in the Ocean?", *PLoS Biology*, 9(8): e1001127.

Morgan, M. G. and Keith, D. W., 1995. “Subjective Judgments by Climate Experts”, *Environmental Science & Technology*, 29(10): 468A-476A.
Porter, T. M., 1995, *Trust in Numbers. The Pursuit of Objectivity in Science and Public Life*. Princeton: Princeton University Press.
Refsgaard, J-C., van der Sluijs, J. P., Brown, J. D., and van der Keur, P., 2006. “A framework for dealing with uncertainty due to model structure error”, *Advances in Water Resources*, 29(11): 1586-1597.
Revelle, R. and Suess, H. E., 1957. “Carbon Dioxide Exchange Between Atmosphere and Ocean and the Question of an Increase of Atmospheric CO_2 during the Past Decades”, *Tellus* 9(1): 18-27.
RIVM, 1996. *Milieubalans 1995*. Alphen aan den Rijn: Samsom.
RIVM, 1997. *Milieubalans 1996*. Alphen aan den Rijn: Samsom.
RIVM, 1998. *Milieubalans 1997*. Alphen aan den Rijn: Samsom.
RIVM, 1999. *Milieubalans 1998*. Alphen aan den Rijn: Samsom.
RIVM, 2000. *Milieubalans 1999*. Alphen aan den Rijn: Samsom.
RIVM, 2001. *Milieubalans 2000*. Alphen aan den Rijn: Samsom.
RIVM, 2002. *Milieubalans 2001*. Alphen aan den Rijn: Kluwer.
Saltelli, A., Guimarães Pereira, Â., Van der Sluijs, J.P. and Funtowicz, S., 2013, “What do I make of your latinorum? Sensitivity auditing of mathematical modelling”, *International Journal of Foresight and Innovation Policy*, (9), 2/3/4, 213–234.
Saltelli, A., Stark, P.B., Becker, W., and Stano, P., 2015, “Climate Models As Economic Guides Scientific Challenge or Quixotic Quest?”, *Issues in Science and Technology*, 31(3).
Taylor, B. N., Parker, W. H. and Langenberg, D. N., 1969. *The fundamental constants and quantum electro-dynamics*. New York: Academic Press.
Thomas, C. D., Cameron, A., Green R. E. Bakkenes, M., Beaumont, L. J. Collingham, Y. C., Erasmus, B. F. N., Ferreira de Siqueira, M. Grainger, A., Hannah, L., Hughes, L., Huntley, B., Van Jaarsveld, A. S., Midgley, G. F., Miles, L., Ortega-Huerta, M. A., Peterson, A. T., Phillips, O. L., and Williams, S. E., 2004. “Extinction risk from climate change”, *Nature*, 427: 145-148.
Urban, M. C., 2015. “Accelerating extinction risk from climate change”, *Science*, 348: 571-573.

Van der Sluijs, J. P., 2002. "A way out of the credibility crisis around model-use in Integrated Environmental Assessment". *Futures*, 34: 133-146.

Van der Sluijs, J. P., 2005. "Uncertainty as a monster in the science policy interface: four coping strategies", *Water science and technology*, 52(6): 87–92.

Van der Sluijs, J. P., 2012. "Uncertainty and dissent in climate risk assessment, a post-normal perspective", *Nature and Culture*, 7(2): 174-195.

Van der Sluijs, J. P., van Eijndhoven, J. C. M., Shackley, S. and Wynne, B., 1998. "Anchoring Devices in Science For Policy: The Case of Consensus Around Climate Sensitivity", *Social Studies of Science*, 28: 291-323.

Van der Sluijs, J. P., Risbey, J. and Ravetz, J. (2005b). "Uncertainty Assessment of VOC emissions from Paint in the Netherlands", *Environmental Monitoring and Assessment*, 105: 229-259.

Van der Sluijs, J. P., Petersen, A. C., Janssen, P. H. M., Risbey, J. S. and Ravetz, J. R., 2008. "Exploring the quality of evidence for complex and contested policy decisions", *Environmental Research Letters*, 3, 024008 (9pp).

Van der Sluijs, J. P., van Est, R. and Riphagen, M., 2010. "Beyond consensus: reflections from a democratic perspective on the interaction between climate politics and science", *Current Opinion in Environmental Sustainability*, 2(5-6): 409–415.

Van der Sluijs, J.P., Wardekker, J. A., 2015, "Critical appraisal of assumptions in chains of model calculations used to project local climate impacts for adaptation decision support - The case of Baakse Beek", *Environmental Research Letters*, 10(4): 045005.

Zimmer, C., 2015. "Study Finds Climate Change as Threat to 1 in 6 Species", *New York Times*, 30 April. Available at: http://www.nytimes.com/2015/05/05/science/new-estimates-for-extinctions-global-warming-could-cause.html