3 Philosophical Issues in Model Assessment

NAOMI ORESKES[1] AND KENNETH BELITZ[2]
[1]*Department of History and Program in Science Studies, University of California San Diego, USA*
[2]*Water Resources Division, US Geological Survey, USA*

3.1 Introduction

In recent years, there has been an explosive increase in the use of models in a wide variety of fields – pharmacology, toxicology, economics, forest ecology, climatology and psychology are some examples. With increased use has come increased concern about model assessment. How can we tell if a model of a complex system is a good model? How can we judge the relative strengths of different models? How can we test a model that we wish to use in a predictive mode? In this chapter, we consider these issues in light of examples from both hydrology and other disciplines in which similar issues comes to the fore.

In many disciplines model results can be compared with independent lines of evidence. In pharmacology, a simulation model can be compared with a clinical trial. In aeronautics, simulations can be compared with physical results in a wind tunnel or from a prototype. In chemistry, numerical output can be compared with laboratory experiments. In hydrology, however, the geographical and chronological scales of processes and the difficulty of access to the subsurface make it difficult to find a suitably comprehensive independent line of evidence against which to test model claims. So how do we judge the reliability of the knowledge the model provides? Both the makers and the users of models want to know whether a model accurately reflects the natural processes it claims to represent.

The inherent uncertainties of models have been widely recognised, and it is now commonly acknowledged that the term 'validation' is an unfortunate one, because its root – valid – implies a legitimacy that we are not justified in asserting (Tsang 1991, 1992; Anderson and Woessner 1992; Konikow 1992; Konikow and Bredehoeft 1992; Oreskes et al. 1994; Beck et al. 1997; Oreskes 1998; Steefel and van Cappellen 1998). But old habits die hard and the term persists. In formal documents of major national and international agencies that sponsor modelling efforts, and in the work of many modellers, 'validation' is still widely used in ways that assert or imply assurance that the model accurately reflects the underlying natural processes, and therefore provides a reliable basis for decision-making. This usage is misleading and should be changed. Models cannot be validated. The reasons why have been outlined in detail elsewhere (Konikow and Bredehoeft 1992; Oreskes et al. 1994). Here, we summarise these reasons before proceeding to a more detailed discussion of conceptual issues that pose problems even for models that seem to work.

Model Validation: Perspectives in Hydrological Science. Edited by M.G. Anderson and P.D. Bates.

It is widely acknowledged that all scientific knowledge is provisional. It has to be. If scientific enquiry is a process of discovery of new knowledge and refutation of old mistakes, then the knowledge we have at any given moment must be provisional. For science to advance, we must be critical of existing knowledge. Therefore all scientific knowledge is, in some sense, uncertain, and the issues of uncertainty inherent in model validation are not unique to modelling. However, there are particular aspects of numerical simulation models that exacerbate the problem of uncertainty to a degree that may be substantially greater than in some forms of scientific endeavour. These features are non-uniqueness, the problem of temporal and spatial divergence, and the subjectivity of model assessment.

1. *Non-uniqueness.* A fundamental issue in modelling is the problem of non-uniqueness: more than one model configuration may produce the same output. The more complex the model, the more opportunity for alternative model realisations.

   There are three major forms of non-uniqueness: numerical, parametrical and conceptual. Bethke (1992) has emphasised numerical non-uniqueness: the possibility of more than one solution to the governing equations. Even a well-constrained model may give a uncertain result if the governing equations can be solved in more than one way. Konikow and Bredehoeft (1992) have emphasised parametrical non-uniqueness: there can be a wide range of possible model inputs to achieve an expected output. The more complex the problem being addressed, the greater the likelihood of significant non-uniqueness in the model solution. Below, we discuss the problem of conceptual non-uniqueness: that more than one conceptual model may prove adequate to account for the empirical evidence. Because most problems in the earth and environmental sciences are inverse problems – we know the configuration of the world, but we lack knowledge of the processes and parameters that produced it – we always face the problem of non-uniqueness.
2. *Temporal and spatial divergence.* Models are frequently calibrated or tested against historical data for a given region or time frame before using them to predict the behaviour of the system. Typically, however, our calibration or history match is smaller or shorter than the space or time frame that we want to predict. Although the model may accurately reproduce available observational data, there is no guarantee that it will perform at an equal level when applied over a larger geographic scale or a longer time frame. There may be small errors that do not impact the model fit, but which, when extrapolated over much longer time frames or much larger geographical areas, will generate significant deviations.

   A good example of this is given by Konikow and Bredehoeft (1992) in an analysis of a 40-hour pump-test in the Dakota sandstone, one of the most important aquifers in the continental United States. The results of the pump-test were consistent with two models: the Theis solution, which assumes no flow through confining layers, or the Hantush solution, which allows for transient flow through confining layers. Konikow and Bredehoeft estimate that the pump-test would have to be run for more than 1000 years for a divergence to be detected. Yet some models calibrated over short time frames attempt to predict regional flow over geological time frames. A model may also diverge from historical data because the system it describes changes – this point is discussed at length below.
3. *Subjectivity of model assessment.* The possibility of magnification of small errors over time and space leads to another source of uncertainty in modelling: the subjectivity of our judgements of what constitutes a 'good' fit. The literature of model validation is filled with terms like 'adequate', 'acceptable', 'satisfactory', and 'reasonable', even in sophisticated mathematical treatments. These are obviously subjective terms and they highlight what should be an obvious point: all models are approximations. Most modellers freely acknowledge this. But how do we determine what constitutes a good enough approximation?

> Because the definition of good is inherently subjective, there is unavoidable uncertainty associated with the definition of acceptable error. As put by Konikow and Bredehoeft (1992), 'under the common operational definitions of validation, one competent and reasonable scientist may declare a model validated while another may use the same data to demonstrate that the model is invalid'. There is no way to eliminate subjectivity and value judgement when we ask ourselves what constitutes a 'reasonable' degree of uncertainty (Bredehoeft and Konikow 1992; Konikow 1992).

For these reasons, we argue here, as we have elsewhere, that the language of validation is unhelpful and should be avoided (Oreskes et al. 1994). We should prefer neutral terms like evaluation or assessment. Good science requires a critical stance, but the language of validation shifts our focus in a different direction. It implies that we can provide firm assurance of the reliability of our models. In fact, the most we can do is to say that a model is close to the state of the art (if it is), that it has been grounded in our best understanding of known natural processes (if it has), and that we built it on the basis of abundant, well-constrained empirical input (if we did).

Even if we use a neutral term like model evaluation, problems remain. How *do* we evaluate a complex model? How do we differentiate between better and worse models? What should we do if we need to use a model in a predictive capacity for social, political or economic purposes? These questions motivate the remainder of this chapter.

3.2 THE PROBLEM OF PREDICTION

Much of the demand for model assessment derives from the desire to use models in a predictive mode. A large number of models have been built in response to environmental problems – nuclear waste disposal, acid rain, water supply, groundwater contamination – which involve forecasting the behaviour of complex natural systems (Sarewitz et al. 2000). Models that involve predictive capacity raise a particular set of issues related to a basic fact: predictions are generally wrong. We might put this another way: predictions are always wrong, in the sense that we can't and don't expect any model to be precisely correct in all respects. What is more significant is that models are often wrong in systematic ways. This is true in most areas of human endeavour and the earth sciences are no exception. While the number of studies that examine the predictive accuracy of model results is modest compared to the number of models that have been built, the available studies are clear in their results. We do not know enough about complex natural (or human) systems to be able to predict them reliably. This leads to two questions: Why are we making predictions? What can we learn from past mistakes?

3.2.1 Are Predictions Necessary?

If predictions are unreliable, then why do scientists make them? When models are built in aid of public policy, scientists may feel that they have to make predictions to serve the agencies and constituencies that support them. While there may be good reasons to run a model in a predictive mode, it is important to realise that public policy does not necessarily require it. In many cases it is possible to develop sound public policy based on a general scientific understanding without specific scientific predictions (Brunner and Ascher 1992; Sarewitz and Pielke 1999; Rayner 2000). It is important for scientists to understand this so as not to feel compelled to make (bad) predictions simply because there seems to be no other choice.

A good example is global climate change. A great deal of climate research is dedicated to the construction of complex General Circulation Models (GCMs) with the goal of predicting the

effects of increased atmospheric carbon dioxide. This is in part the result of a policy decision. In the early 1990s, the first Bush administration in the United States defined the primary problem of climate change to be scientific uncertainty, and made research its central policy response (Brunner 1991; Brunner and Ascher 1992; Rayner 2000). Eliminating or at least greatly reducing uncertainty was considered a prerequisite for action, and this led to greatly increased funding for climate research, particularly climate modelling. Scientists did not object; the idea of knowledge as a basis for action is eminently reasonable. And what scientist would protest better funding? Yet, despite (or perhaps because of) recent advances in understanding, it is clear that significant reductions in predictive uncertainty will be some time in coming. If global warming is real, and if it is happening now, then we may not be able to wait for predictive accuracy. By the time we have achieved it, damage will have been done and may be impossible to undo.

In response to external pressure to generate predictions, one can point to the existence of alternatives strategies and complementary courses of action. In the case of climate change, for example, society could focus on mitigation and adaptation without precise predictions. As Rayner (2000) points out, steps can be taken to protect and increase human welfare that would be beneficial to the populations involved even if the predicted climate change does not occur. Policy analysts refer to this as a 'no regrets' strategy – to act so people will benefit even if the feared event does not occur. There are positive local benefits to be accrued from reductions in atmospheric pollution (e.g. in human health, protection of ecological habitat, preservation of viewsheds, the pleasure of breathing fresh air) irrespective of whether global warming turns out to be a profound global threat. It can be rational to act despite uncertainty, and actions can be evaluated as one goes along. Finally, even if we had an accurate predictive model of the effect of atmospheric carbon dioxide on global climate, it would not in itself dictate appropriate policy response. Policy decisions hinge as much on the values of communities as they do on the facts of science.

Scientists can also create alternative avenues of study. An example is earthquake prediction. In the 1970s, hopes were high for accurate forecasting of earthquakes in California within a few years, and some seismologists went so far as to claim that short-term warning of impending earthquakes was imminent (Nigg 2000). Today few if any scientists consider this a realistic goal. The failure of earthquake prediction has led to a shift in scientific focus toward the study of seismic wave propagation and material response, which in the long run may prove efficacious in reducing seismic hazard. Meanwhile, policy-makers have focused on societal preparedness: emergency response plans, back-up medical facilities, individual family preparedness etc. (Nigg 2000). In hindsight one can see that accurate prediction of individual earthquakes may be less important to society at large than a general appreciation of seismic risk and an appropriate overall pattern of planning and preparedness.

Given the difficulty of making accurate predictions and the insight that public policy may not necessarily require them, it seems clear that model assessment need not focus solely on predictive capacity. One can gain insight and test intuitions through modelling without making predictions. One can use models to help identify questions that have scientific answers.

One approach that has been developed recently is to acknowledge that available data is often insufficient to select one model uniquely among many that might be built. Keith Beven and colleagues refer to this problem as equifinality – a form of non-uniqueness – and developed the GLUE procedure (Generalised Likelihood Uncertainty Estimation) to address it (Beven 1993, 1996; Aronica et al. 1998; Hankin and Beven 1998a, 1998b). The purpose is to identify the range of model output that is generated from a range of feasible model input. Gupta et al. (1998) also address the issue of model non-uniqueness. They develop a multi-objective optimisation procedure for identifying families of solutions that satisfy various combinations of different measures of model performance.

These efforts help to focus attention away from expectations of certainty, and underscore the fact that models of natural systems are simplifications of complex systems. By formally recognising a range of possible model input and corresponding output, these methods alert policy-makers and the public that model output is not the same as an accurate prediction.

But what happens when scientists are asked to make predictions to support public policy (Beck et al. 1997)? Whether one makes a single prediction, or generates a range of predictions, there are still conceptual issues at stake. Reality may turn out to lie beyond the range of model prediction. Therefore it is worthwhile to understand how and why model predictions have gone wrong in past. The errors in models are frequently non-random. In fact, they appear to be systematic in particular and recognisable ways.

3.3 SYSTEMATIC ERROR AND BIAS

The incompleteness of our knowledge of natural systems opens the door to systematic error and bias. We can think of model input as falling into three categories: factors that are known and measured, factors that can be estimated based on informed judgement (e.g. based on prior experience in systems that are believed to be similar), and guesswork. All models involve informed judgement and typically a bit of guesswork as well. Wherever subjective judgements are required, the potential exists for systematic error and bias.

Given the diversity of human attitudes and opinion, one might hope that individual bias and idiosyncracy would tend to cancel out over the course of modelling efforts. For example, construction of more than one model of a system can reveal biases and errors in a single model. This can be an argument for a 'competing teams approach'. Different modellers or groups of modellers ought to have different subjective tendencies, and the totality of independently constructed models might converge on a correct result. But redundancy in modelling efforts is very costly. Even if we pursue it, it may not solve our problems, because systematic bias in professional communities can cause independent groups and individuals to bias their results in similar ways (Fischhoff 1982; Tversky and Kahneman 1982; Ascher and Overholt 1983; Ascher 1981, 1993).

Ascher (1993) presents a *post hoc* analysis of 21 development projects funded by the World Bank, all of which had been subject to rigorous *ex ante*, quantitative rate-of-return analyses. In each case, the analyses were performed by educated, experienced professionals, working under conditions in which the criteria for assessment were codified. Ideally such conditions should leave relatively little room for subjective bias, or at least render any such bias individual and random. But not so. The results of these analyses were systematically and seriously biased toward overly optimistic results. The true rates of return on these projects were substantially lower than predicted, and many projects were approved which in retrospect should not have been, given the evaluation criteria.

In a related study, Isham and Kaufmann (1995) analysed *ex ante* appraisals for over 1200 World Bank projects, in which the average predicted rate of return was 22%. The actual average rate of return turned out to be 12%, just skimming the 10–12% threshold value set by the World Bank for approval. Many individual projects fell well below this threshold value and several had a negative rate of return. Perhaps more important, many of these projects caused substantial environmental damage. These detrimental effects were not wholly unanticipated, but they had been justified *ex ante* on the grounds of the expected economic benefits, benefits that may have made alternative patterns of development seem ineffectual. In retrospect, however, the exaggerated economic benefit appears as a significant factor in discouraging alternatives.

Why were these analyses systematically biased, and what relevance does this example have for

hydrology? One answer is human nature. While the human factors may seem obvious, they bear repeating because they apply to natural scientists as well as to social scientists: positive appraisals create work for the appraisers. The World Bank is in the business of lending money for development projects, and an appraiser who consistently rejected proposed projects would soon find him or herself under pressure. There is a parallel here with models in the natural sciences: much of the funding for modelling comes from agencies that want to act on a specific public policy or environmental problem. And who is more likely to be funded: The modeller who says he or she can produce the desired predictions, or the one who says he or she can't?

The appraisers of economic models generally share the goals of those proposing the projects being modelled: they believe in economic development and are therefore willing to provide analyses despite incomplete information. Again there is a parallel: most scientists believe that scientific information can and should form the basis of rational public policy. Scientists share the goals of agencies that want to use the available scientific knowledge as the basis for action, and this makes them susceptible to overestimating the capability of the model to provide such a basis.

Shared goals and commitment to scientific solutions may help to account for why modellers construct models with incomplete information, but do not explain why the models are biased in a particular direction. There are obvious political reasons why promoters of projects would wish to exaggerate their potential benefits, but the analysts at the World Bank are professionals who, like natural scientists, consider it their job to make informed and objective judgements based on quantitative analysis. A professional who consistently made inaccurate forecasts would seem to be placing his or her credibility at risk. Why would seasoned professionals consistently overestimate the benefits of proposed development projects? Ascher (1993) argues that two technical factors contributed to systematic error in this case: (i) inadequate incorporation of negative impacts of unknown or unlikely effects and (ii) an optimistic bias with respect to implicit conditionals.

3.3.1 Inadequate Incorporation of Unknown or Unlikely Events

All models involve parameters or processes that are unknown or poorly known, and many models involve the problem of estimating unlikely events. What do we do about it? In the case of parameters that are poorly known, modellers may take a 'best-guess' approach. Lacking quantitative data on a parameter, they make a best assessment of what it might be. Typically this is done on the basis of past experience in systems believed to be similar. In the case of parameters that are unknown, modellers may leave them out.

Ideally, we would improve our knowledge of the input parameters, and a large number of researchers are actively doing this. The methods typically draw upon relationships between model input parameters and auxiliary data, or utilise mathematical inverse methods to estimate the input parameters based on limited measurements of those values and more extensive state variables (e.g. Hill 1992). These methods improve capacity to specify model input. Nevertheless, in many modelling problems, input parameters remain incompletely known, and there may still be unknown effects.

By definition, an unknown effect cannot be incorporated into a model. But if an unrecognised factor rears its head, then the model will be in error. In highly structured, settled human societies, unanticipated events are almost always costly: they require responses (which cost money), and they may cost lives. Leaving out unknowns in models that involve human systems almost always biases a model in an optimistic direction (Ascher 1993). Reality is likely to be worse than we wish.

There is a similar pattern with low probability events. The rarer an event is, the more difficult

it is to study and analyse. While considerable progress has been made in the statistical analysis of rare events, it remains a vexing domain. Many rare events leave little or no trace, and therefore are almost impossible to analyse. If you cannot analyse something, then you cannot generate a meaningful quantitative measure of it for model input. If an event has never occurred, or never occurred in this particular context, it is impossible to quantify its likelihood.

Scientists are understandably loath to make up values where none exists. Lacking adequate basis for analysis, modellers may leave out extremely low probability events. The small probability of the event discourages both its analysis and its incorporation. Yet rare events do occur, and largely because they are unanticipated their impact is generally negative. The omission of very low probability events exaggerates our capacity to make accurate predictions and biases our models in an optimistic way.

3.3.2 Optimistic Bias in Implicit Conditionals

Ascher points out that a model is a conditional proposition: if certain factors are in place, other things will follow (Ascher and Overholt 1983; Ascher 1993). Yet the proposed factors are almost never in place in exactly the way that the model proposes, and the effect of deviation is generally negative. This is analogous to the phenomenon of construction delays, well known to anyone who has renovated a kitchen or awaited an office in a new university building: there are many factors that make projects take longer but few that speed them up (Ascher 1993; see also Tversky and Kahneman 1982, p. 16; Gutierrez and Kouvelis 1991).

The idea of a model as an implicit conditional applies to natural systems as well as human behaviour, because running a model in a predictive mode presupposes that the driving forces of the system will remain within the bounds of the model conceptualisation. If the driving forces change, then even a well-calibrated model will fail. Consider the following examples.

Example One: The Salt River and Lower Santa Cruz River Basins

Konikow (1986) presents the example of the Salt River and lower Santa Cruz River basins in Central Arizona, an arid area of intense groundwater use. In the mid-1960s, a groundwater model was built and calibrated on the basis of 41 years of historical data on pumping and water levels (1923–1964), and used to predict future water level changes for the next ten years. On the basis of the model's success in matching the historical data, the modeller had concluded that the model was a valid representation of the system and therefore could 'be used to predict future ground-water conditions' (Konikow 1986, quoting Anderson 1968; see also Konikow and Patten 1985). In the early 1980s, the predictions were compared with what really happened. The model predictions were systematically wrong; real water levels were consistently higher than predicted. Groundwater levels had been falling throughout the calibration period due to groundwater pumping, and the model predicted that they would continue to fall. Major declines were predicted everywhere, but the real changes were either smaller than anticipated or in the opposite direction.

Why did the model fail? In retrospect, it can be seen that the system changed almost as soon as the model was produced. The 20-year period prior to model construction was one of relatively uniform downward trend in water levels changes driven by a consistent upward trend in groundwater pumping. But, as Konikow explains, 'a marked break in this trend occurred very soon after the end of the calibration period'. In the mid 1960s, people began to pump less. They may have done so because costs were rising as the water table was falling; because farmers took steps to increase irrigation efficiency in response to rising costs; because more surface water was made available; or because land was taken out of production in response to rising costs or

encroaching urbanisation. It is unclear which of these factors was most significant, but it is clear than human factors played a large role in the inaccuracy of the predictions. Although this was a model of a physical system, human activity was decisive in undermining its predictive capacity.

Although it is not news that humans are unpredictable, many models in the natural sciences implicitly assume consistency in human behaviour. Few terrestrial systems are completely closed to human effects, and this is particularly so for the systems we are likely to be modelling. If the purpose of a model is to aid in decision-making, by definition this means that the system is of interest to humans and very likely impinged upon by human behaviour. But it goes against the grain of physical scientists to acknowledge this. We have been trained to study 'natural' systems, and only recently have we begun to assimilate the fact that there is no sharp line between human and 'natural' systems.

Beyond demonstrating the role of human factors in models of natural systems, there are two other striking features of the Salt River study. The first is that modellers may have insights that are not incorporated into their models. In this case, the modeller noted that the assumption of consistency in driving forces was unlikely: actual pumping rates would probably fall as costs continued to rise. If this happened, then the predicted declines might be exaggerated – and they were. (In hindsight, the model stands as a marker of the changes that were imminent: the situation was under pressure, people were concerned – which is why the model was built – and changes in human behaviour were likely.) But although the modeller acknowledged this in discussions of the model, he did not incorporate this insight into the model. To predict a change in the driving forces of a system may require explicit insertion of the modeller's judgement; the goal of objectivity makes modellers reluctant to do this. The result is models that fail to encompass the full extent of their builders' insight and intuition.

The second striking feature of this study is that the effects of human behaviour are not the whole story. Konikow's analysis revealed that the spatial pattern of predictive error was not explained by changes in groundwater pumping. The correlation between the spatial distribution of error in assumed pumping and the spatial distribution of error in water levels was very low (-0.086). Again, these errors were not random. Certain areas were much worse than others, and errors in nearby wells tended to be close in value. Yet Konikow was unable to correlate the pattern of error with any known factor in the model. Something else in the model was wrong, and even retrospective analysis failed to reveal what it was. This leads to an important conclusion: most complex models probably contain more than one source of error (see also Alley and Emery 1986). If so, then the various errors may cancel each other out to create a model that fits the historical data, but is still flawed. The more complex the model, the greater is the prospect for compensating errors, and the more likely it is that such errors will be undetected.

Example Two: The Cochella Valley

A second example comes from the Cochella Valley region in Riverside County, California, east of Los Angeles (Konikow and Swain 1990; Konikow and Bredehoeft 1992). Here a model was built to predict the effect of an artificial recharge programme on groundwater levels and dissolved solute concentrations. The model was calibrated on the basis of nearly 40 years of historical data (1936–1973), and used to predict water levels over the next seven years. As in the example described above, the modeller presumed that the success of the model in reproducing historical data was evidence that the model accurately represented the system, and could be used 'to predict water level changes from projected pumpage and (or) projected artificial recharge' (Konikow and Swain 1990, quoting Swain 1978).

The model predicted that groundwater levels would continue to drop despite the effect of the recharge programme. Again, the results were systematically in error. *Post hoc* comparison of 92

wells showed that the magnitude of the decline was consistently over-predicted. For individual wells, there was a wide range of error. Errors were greatest in and near tributary canyons entering the main valley. In hindsight, the period 1974–1980 was unusually wet, leading to high levels of recharge from creeks discharging from the tributary canyons.

The Cochella Valley case is similar to the Salt River case: the model was in error because of changes in the forcing function of the system, in this case climate. The model calibration implicitly assumed that the previous 40-year record covered the full range of conditions that operate in the valley, but this turned out not to be the case. The ability of the model to reproduce historical data did not translate into a capacity to predict future conditions.

3.4 LESSONS FROM EXPERIENCE?

These examples highlight a number of points. When model parameters are adjusted to obtain a best fit with historical data, it introduces a bias towards extrapolating existing trends. If those trends do not persist, the model is likely to be in error. Calibrated models are biased in favour of stasis.

The bias of stasis may operate in two ways, one explicit and one implicit. When a model assumes a particular driving force, like climate or pumping rates, then the extrapolation of that driving force into the future explicitly embeds an assumption of stasis into the model. If the driving forces change, then the model will diverge from actual conditions (Konikow and Patten 1985; see also Alley and Emery 1986). Stasis may also be implicitly embedded. A model will diverge from reality if the data gathered in the calibration period were in some way unrepresentative, even if the overall driving forces of the system remain the same. The process of calibration against historical data assumes the representivity of that data, and therefore embeds an assumption that the conditions they represent are on-going (Konikow and Person 1985). Finally, as noted above, a calibrated model may contain more than one set of errors, which cancel each other out. If conditions remain unchanged, the errors may continue to cancel each other out, and the model continues to work. But if operating conditions change, the errors may no longer cancel each other out, and the model begins to break down.

These examples show that our models may be less nuanced than our subjective understanding of the natural systems we are modelling. Scientists may have insights about systems that they are reluctant or unable to incorporate into their models. The goal of objectivity and adherence to Ockham's razor (the principle that we should prefer simpler explanations) create resistance to incorporation of nuance and subjective judgement. Yet it may be precisely within the realm of nuance and subjective judgement that our best understanding of a natural system lies.

To use William Ascher's terminology, modellers may recognise that their models are implicit conditionals – the predictions will come true if and only if there are no major changes in the forcing functions of the systems – but this insight is not often articulated, and the users of the model may be unaware of it. If they are aware, they may still choose to ignore it in order to proceed with the task at hand. The modeller may later be blamed for faulty predictions caused by limitations of which he or she was well aware, and perhaps even warned people. An obvious point here is the value of scenario development. The Salt River and lower Santa Cruz River model could have been run using a range of assumptions about future pumping rates or possible human responses to greater pumping lifts; the Cochella Valley model could have been run with 'extreme' case values for rainfall based on geological and meteorological evidence and insight rather than temporally limited historical records.

3.5 REPRODUCING THE BEHAVIOUR OF A SYSTEM VERSUS CAPTURING CAUSAL PROCESSES

The examples presented above involve retrospective analysis, but how should we evaluate a model in the present? Many modellers believe that the capacity of a model to reproduce data from the natural world – past or present – is evidence that it will be able to reproduce future behaviour. De Marsily and colleagues (1992) write, for example: 'We all know that the parameters of a model are uncertain, probably wrong in many cases, and can easily be invalidated. . . . So what? As long as they reproduce the observed behavior of the system, we can use them to make predictions.' Model post-audits refute this assertion. The examples described above successfully reproduced the observed behaviour of the modelled systems, yet failed to make accurate predictions.

The capacity to reproduce a response may or may not be sufficient grounds for prediction. Consider a proposed mine in which one needs to anticipate the day-to-day variability of the ore feed into the mill. A geostatistical model of the ore deposit can be developed that accurately represents the variability of the ore to help guide the design of the mill. The cause of the variability is irrelevant, and the model provides an adequate basis to design the mill. Now consider an exploration team, working at the same site, trying to find more ore. For this purpose, understanding the cause of the underlying variability – the geological controls on ore grade – is essential, and the model is useless for predicting where to look. In this case, no one would claim that the model addresses the cause of ore grade, and no one would make the mistake of imagining it could be used in aid of exploration. Yet there are many cases where people do make precisely this mistake: they presume that because a model accurately mimics the behaviour of a system, it must encompass an accurate representation of the underlying causal processes.

Houser et al. (1998) present a method for generating state variables, such as moisture content, which explicitly merges simulation-based estimates and observational data. As in the case of geostatistical methods, if the goal is the generation of a distributed parameter on a grid, then the method provides an efficacious means of doing that. However, the method does not seek to explain the reasons for the divergence between the simulation-based estimate and the observational data.

To generalise: if the underlying causal processes are irrelevant to the task at hand, then a calibrated model may make perfectly accurate predictions and prove highly reliable. One can design a mill without understanding the causes of ore grade variability; one can design a dam without worrying about the causes of rainfall fluctuation. But if the underlying processes are relevant – as in the case of the Salt River or Cochella Valley models – and the model fails to capture them, then the model is likely at some point to fail (see also Klemes 1982; Konikow and Patten 1985; Bredehoeft and Konikow 1992).

Validation is sometimes described as a process of confidence-building (de Marsily et al. 1992; Neumann 1992), but the most common approach to confidence building – reproduction of historical data – may be no more than mimicry. The capacity to mimic data is not evidence that you have captured underlying processes, and therefore not evidence of predictive capacity.

3.6 CONCEPTUAL ERROR

Konikow and Bredehoeft (1992) point out that the Cochella Valley example can be viewed in two ways: (i) that the historical database was not long enough, because it did not cover the full range of natural conditions, or (ii) that the model involved a flawed conceptualisation, because it

inadequately appraised the highly variable nature of recharge in the desert environment. From the second perspective, the problem is not that the historical data were insufficient, but that the model conceptualisation was wrong. It embedded a faulty assumption: that the near future would closely resemble the near past.

Conceptualisation is probably the most thorny issue in modelling. It is the foundation of any model, and everyone knows that a faulty foundation will produce a faulty structure (Tsang 1991; Anderson and Woessner 1992; Konikow and Bredehoeft 1992). Yet what to do about it remains a problem. Much attention in model assessment has focused on quantification of error, but how does one quantify the error in a mistaken idea? Some refer to this as the epistemological uncertainty, to emphasise its foundational aspects (Funtowicz and Ravetz 1985, as cited in Rayner 2000). It is uncertainty rooted in the foundations of our knowledge, a function of our limited access to and understanding of the natural world. Almost by definition, conceptual error cannot be quantified. We don't know what we don't know, and we can't measure errors that we don't know we've made. In most cases, it is difficult to identify our errors at all. Not being able to identify them, we hope that they do not matter, but they may, because faulty conceptual models can lead to faulty conclusions and wrong courses of action. The following examples illustrate the point.

3.6.1 Conceptual Error, Example 1: Isostasy and Continental Drift

An example from the history of the earth sciences shows how scientists can come to radically wrong conclusions on the basis of a faulty conceptual model (Oreskes 1999).

In the early 20th century, there was a spirited debate over Alfred Wegener's theory of continental drift, which proposed that the earth's continents were mobile, and had been rearranged in various configurations during the long course of earth history. Wegener's theory was a reasoned response to a fundamental problem in the earth sciences: how to explain the palaeontological evidence that plants and animals had once freely migrated between what are now widely separated continents. In the late 19th century, this question was answered by the theory of sunken continents: that early in earth history, the entire surface of the earth was covered by a giant supercontinent, Gondwanaland, whose pieces collapsed to form the ocean basins. Organisms that had once freely migrated were subsequently isolated.

Most geologists accepted this explanation until it was refuted by the proof of isostasy: the idea that the continents sit in hydrostatic equilibrium within a denser substrate. In the early 20th century, studies of regional variations in gravitational acceleration proved that the continents are composed of less dense material than the ocean floors, and float within them. If so, then the theory of sunken continents was flawed. Wegener proposed drifting continents as an alternative explanation of the palaeontological data that did not conflict with isostasy.

Because Wegener's theory was premised on the principle of isostasy, one might expect that geodesists would have been among his strongest supporters, but in fact they were among his most vocal opponents, convincing many geologists to reject the idea despite its obvious explanatory power. Why? There were in fact two competing conceptualisations of isostasy. One, called the Airy model, supposed that mountains float as icebergs do, with large, hidden roots beneath them, and the depth at which isostatic equilibrium is achieved is proportional to the height of the topography above. The other, called the Pratt model, supposed that mass variations at the surface are compensated by density variations in the rock column below, and the depth at which isostatic equilibrium is achieved is uniform.

The two conceptual models were empirically equivalent. Both were compatible with available geodetic evidence, and they led to the same quantitative predictions with respect to surface gravitational effects. However, the Pratt model, with its assumption of a uniform depth to the

base of the crust, was far easier to use. One could more readily calculate predicted surface effects using the Pratt than the Airy model. In the age before digital computers, this was an important consideration. Its mathematical simplicity was also appealing from the perspective of Ockham's razor. So a consensus formed around Pratt.

The choice of the Pratt model had serious conceptual consequences when continental drift was discussed a few years later, because it implied that there were no large-scale horizontal forces operating in the earth's crust. Most geodesists therefore rejected continental drift. Some became adamant opponents, not merely of the specifics of Wegener's model (which most contemporary scientists would say was flawed), but of the very idea that continents could move in a horizontal fashion – an idea that is now fundamental to earth science.

In later years, on the basis of seismic evidence, the Pratt model was shown to be wrong. Today geophysicists believe that isostatic compensation is primarily achieved by differences in the thickness of the crust, as argued by Airy. But because the Pratt model worked, the scientists who used it came to believe that it was true. From their perspective, their model was 'validated'. It fit the relevant data. It made accurate predictions. It *worked.* But it was not a realistic representation of the earth. Its success at mimicking surface behaviour was not evidence that it had captured the underlying causal structure.

The example of isostasy and continental drift shows how conceptual models can be underdetermined by available evidence, causing an empirically adequate model to break down when applied to a problem other than the one for which it was built. A second example shows how a model based on a faulty premise may fail even at the job for which it was intended.

3.6.2 Conceptual Error, Example 2: The Limits to Growth

In the early 1970s, a group of systems analysts addressed the question we now call 'sustainability'. Their project was sponsored by the Club of Rome, a group of European industrialists, statesmen and scientists concerned about overuse of natural resources. The '*World*' model (very modestly labelled!), described in the widely-read book, *The Limits to Growth*, predicted widespread natural resource shortages, exponential price increases for raw materials, and possibly global economic collapse before the end of the century (Meadows et al. 1972; see also Peccei 1977; Meadows et al. 1992). The modellers concluded that the industrialised nations had to decrease their consumption of natural resources, and fast.

The end of the century has come and gone and resource use continues to grow, but real prices are level or down for virtually all commodities, and reserves of most natural resources are greater today than when the model was built (Hodges 1995; Moore 1995). While there has been economic change, much of this has had to do with falling rather than rising commodity prices (Simon and Kahn 1984; Tierney 1992).

One reason why the predictions of the *World* model have not come true is obvious in hindsight: the static way in which the model treated natural resource commodities. Resources such as copper, chromium, silver and gold were treated in the model as fixed and finite masses whose volume could only decrease over time. The greater the use rate, the faster the depletion. This seems like common sense; many people cannot imagine how it could be otherwise. On one level it is indisputable: the mass of any element in the earth is a finite (albeit unknown) number. However, from the perspective of resources management and sustainability, this view is inadequate because the resource of something is not the same as the mass of it in the earth. By definition, a resource is something that may be used by humans (Hodges 1995; Moore 1995). This involves many variables, including the capital and technology available to find and extract it, the cost of labour, and the price people are willing to pay for it. Proven reserves are an even more constricted thing: they are only that portion of a resource that has been discovered,

delineated and measured, and can be exploited under given conditions (Hodges 1995; Moore 1995; see also Oreskes 1998).

The *World* modellers made an elision between the proven reserves of a metal and its total mass in the earth as if they were the same thing. But they are not the same. Over time, the total mass of a metal in the earth must decrease or stay the same, but proven reserves can increase as a result of increased exploration, improved technology and/or decreased costs. Proven reserves of most metals *have* increased since the early 1970s, primarily because of more effective geological exploration during the subsequent decades; prices have fallen as a result (Hodges 1995).

This is not to say that the world will never run out of resources, that commodity prices will never rise, or that there are not important social, political, environmental and economic dimensions to sustainability. But it does illustrate the way in which faulty conceptual models can lead to faulty predictions. What seemed like common sense – increasing use must lead to diminishing stocks – was in error, because the definition of what constituted a 'stock' was misconceived.

The example of the *World* model also shows that a faulty conceptual model can inhibit consideration of alternative courses of action. If natural resources were fixed in the way the modellers conceived, then the industrialised nations would have had no choice but to decrease consumption in order to avoid economic crisis. But if we recognise that resource stocks are a more fluid thing, then other courses of action, such as improved technological capacity for resource recovery, or substitution, come to mind. Even if reserves *were* fixed – even if our technological capacities were utterly static – the use of proven reserves as measure of the availability of a commodity presupposes that the world is fully explored. But it isn't. The world may be known in a geographical sense – 'terra incognita' no longer appears on our maps – but from a geological point of view the world is quite unexplored. There are substantial areas that have scarcely been examined from a mineral exploration perspective, still more that have not been explored with modern technology. One need only to consider the recent discoveries in Russia, China, South America, Australia, and even Europe and North America to see this.

The assumption that the world was more explored than it really was is an example of the general problem of overconfidence in our knowledge. Studies in the social sciences have shown that most people, including experts, tend to overestimate their knowledge and are willing to make predictions even when the information supplied is inadequate to the task. This has been called the 'illusion of validity' (Tversky and Kahneman 1982, p. 9). Remarkably, this illusion persists even when the person is aware of the problem (Tversky and Kahneman 1982; see also Ascher 1989). Our cognitive biases lead us to think we know more than we do, and encourage us to act on knowledge that is inadequate or incomplete.

The two examples discussed above involve overconfidence. A third example illustrates the problem of oversimplification.

3.6.3. Conceptual Error, Example 3: Beach Processes and Erosion

Beach erosion is ubiquitous in the United States, primarily because of human efforts to prevent natural beach migration in order to maintain homes, roads and other human constructions in the near-shore area. Left to their natural state, all beaches will migrate, and under present geological and climatic conditions this migration is generally inland. To prevent shoreline retreat, communities along both the Atlantic and Pacific Coasts of the United States have constructed sea walls, jetties, groins and other forms of 'armouring' designed to stabilise the beach location and prevent sand loss. These efforts generally fail. Evidence suggests that beach-parallel constructions such as sea walls may exacerbate erosion, while beach-

perpendicular constructions such as jetties help trap sand on the up-current side but accelerate erosion on the down-drift side (Hall and Pilkey 1991; see also Dean 1999).

An alternative approach that has become popular in recent years is beach nourishment, also referred to as 'soft stabilization'. In this strategy, lost sand is replaced by mined sand, usually from off-shore locations. In principle, beach nourishment is a direct response to the need for more beach sand, and it could be an environmentally and aesthetically satisfying response (Houston 1991). In practice, it is extremely expensive, and must be periodically repeated. This leads communities considering soft stabilisation to want information on how much it will cost and how long it will last. To do this, groups such as the US Army Corps of Engineers have built models. Commonly, these models produce overly optimistic assessments of the efficacy of this technique. In many documented cases, nourished beaches have lasted far shorter than predicted; in some cases erosion began even before the nourishment process was over (Pilkey 1994, 2000; see also Dean 1999).

Why are the models wrong? Orrin Pilkey and colleagues at Duke University have documented many reasons why models of coastal processes fail to make accurate predictions (Leonard et al. 1990; Pilkey 1990, 1994, 1995, 1997, 2000; Pilkey et al. 1993; Young et al. 1995; cf. Houston 1990, 1991 for counter-arguments). At root the problem is conceptual: the gap between the complexity of the shoreline systems as compared with the simplicity of the models of them.

Most models of shoreline erosion are based on the 'profile of equilibrium' concept, where the shape of the shoreline, described in terms of water depth, is expressed as a simple power law function with two empirical constants: $Y = AX^n$, where Y is the distance offshore, X is the water depth, and A and n are empirical constants. The value used for n is a worldwide average for all beaches, leaving only one parameter as a variable, 'A', which is believed to depend upon sediment characteristics. The shoreline equilibrium profile concept thus effectively asserts that the only factor controlling the profile of a beach is the sand. The complexity of beaches – the wind, weather, sand supply, coastal topography, bathymetry, subsurface geology, organismal action – is reduced to a single equation with a single degree of freedom. One need not be a coastal scientist to suspect that this simplification is an oversimplification, unlikely to account for the observed variability in coastal processes (Pilkey et al. 1993; Pilkey 1994).

Furthermore, as Pilkey and co-workers point out, the very concept of the equilibrium profile is flawed. It assumes that the shoreface is covered with a sand layer that is redistributed by wave action, and that at some depth – the closure depth – the interaction of surface waves with the sea-floor stops. Beyond the closure depth, there is no significant seaward movement of sand. Geologists and oceanographers refute this claim. Bottom currents that transport sand in a seaward direction have been known for a century, and scour channels on the sea floor attest to their widespread efficacy in moving sand beyond the zone of surface effects (Pilkey et al. 1993; Pilkey 1994).

Beach nourishment models also assume an inappropriate driving force. They predict the rate of loss of the artificial beach based on 'normal' or average waves conditions, but it is widely recognised that the bulk of coastal erosion occurs during rare, major storms. This choice is a form of simplification: averages are easier to calculate than extremes. It is notoriously difficult to predict where, when, and how often major storms will occur, much less what the details of their effects will be. The forcing function of coastal erosion – the major storm – is stochastic. Given that this simplification introduces a conceptual error, beach modellers might decide to make probabilistic predictions, as meteorologists do, but this is not generally done. As Pilkey notes, a common goal of modellers is to predict the 10-year lifespan of a nourished beach, but 'we are no closer to predicting the 10-year behavior of a beach than we are to prediction of 10 years of weather' (Pilkey 2000, p. 20).

Given the gross conceptual oversimplification and the use of an inappropriate driving

force, the failure of these models to predict the efficacy of nourishment projects is hardly surprising.

3.7 WHY DO PEOPLE USE 'BAD' MODELS?

In the examples of Pratt isostasy and the *World* model, the modellers had no way to know that their models were flawed, but the models of beach nourishment raise a thornier question: why do people make and use models that they know are conceptually flawed and have perhaps already failed in a predictive mode?

An obvious reason is the political and social context of the decision-making process. In a sense, one may argue that the conceptualisation of beaches as suffering erosion is flawed: beaches left alone do not generally erode, they migrate (Pilkey and Thieler 1992). Erosion is primarily an anthropogenic effect of efforts to stabilise shorelines. An obvious solution to shoreline migration is to remove endangered structures, or let them collapse and build no new ones. A century ago, it was common practice in some areas of the United States to move cottages back from the shoreline (Dean 1999). However, the political and economic power of wealthy beach-front property owners and the presence of large resorts and high-rise hotels that cannot be moved now make yielding manoeuvres difficult.

Since hard stabilisation has been proved to be counter-productive, soft stabilisation is left as the only option when humans refuse to yield. This creates an implicit pressure on modellers to show that this option can work. As in the case of the World Bank, models of beach nourishment projects are commissioned by people who want the job done. So models are built despite past failures, and biased in an optimistic way.

And what constitutes past failure? In one study of 56 large replenishment projects, the US Army Corps of Engineers argued that their models had been successful because the costs of the projects and the amount of sand placed on beaches were close to projections. From an engineering and budgetary perspective, these projects did not fail. However, the study failed to discuss was whether the beaches were actually maintained for the predicted durations. In many cases, the replenished beach had disappeared long before the next replenishment cycle (Pilkey 1995; cf. Hillyer and Stakhiv 1997). The analysts knew what they meant by 'success' – their post-audit was designed to evaluate costs incurred and volumes of materials used – but it was not what would-be beach-goers considered success. It was not what the project was intended to *do*.

The word 'success' – like the word 'validation' – has an ordinary meaning to lay people. If modellers claim that their models were successful, this is generally translated by ordinary people into a claim that the model accurately predicted what would happen in the natural world.

Modellers sometimes argue that their predictions should really be placed in probabilistic terms, but that the decision-makers who fund these projects – members of the US Congress for example – do not understand error bars. Perhaps so, but who will be blamed if the model fails? Not the member of Congress! It is hard to see how it can be in the long-term interest of the scientific community to make predictions that are bound to fail.

Concerns over legal liability also contribute to the use of models that do not work. The frequency of model failure leads to the paradox that there is a positive incentive to use bad models if those models have already been used by others. Because failure is so common, there is pressure to follow precedent to avoid later legal liability (Pilkey 2000). If the government has an established and widely used model, there is an incentive to use it. Standardisation can be a defensive posture.

And often people do not realise or do not remember that previous modelling efforts failed to

make accurate predictions. At least in the United States, there has historically been little funding for long-term monitoring (Houston 1991; Hillyer and Stakhiv 1997). Konikow and Patten (1985) point out that data collection in hydrological systems frequently ends with the study period, making it impossible to compare the model predictions with actual system conditions. (However, this may not necessarily be the case in other countries; see Smith 1990). Many factors combine to provide short-term incentive to build and use models, even if they are likely to fail in the long run.

Finally there is another point, relevant to all modelling, policy – driven or not. That is the view that modelling is cutting-edge science, and therefore inherently good. There are many cases where a scientist will reap more professional benefit from building a flawed model than from doing further empirical investigation. Who is not attracted to the promise of greater rigour in studying messy systems? But if our mathematical and computational prowess exceed our empirical understanding, we may achieve sophistication at the expense of knowledge. We may also achieve it at the expense of the open-mindedness necessary to learn from our mistakes. As Pilkey puts it, the state of the art is not necessarily close to the state of nature (Pilkey 1997; see also Ascher 1978; Ascher and Overholt 1983).

3.8 CONCLUSIONS

The problem of uncertainty is not unique to modelling, but it takes on an added dimension when scientific knowledge is used in support of public policy. Not all models in hydrology are relevant to public policy, but many are, and this makes it all the more important for modellers to be articulate about the sources of uncertainty in their models, and to think about ways to test for hidden errors.

We have raised here a number of issues that contribute to uncertainty in modelling: systematic error and bias; the difference between reproducing the behaviour of a system and capturing the underlying causal processes; and the problems of model conceptualisation. We have tried to highlight some of the ways in which models can go wrong, even when they are calibrated against large databases.

The most difficult issue to address in modelling is conceptual error. The examples provided in this chapter show that models may match available observations, yet still be conceptually flawed. Such models may work in the short run, but later fail. Recent advances in computational power may help us isolate conceptual error: if a set of simulation output, based on exhaustive sampling of parameter space, fails to encompass the observational data, this signals inadequacy in the underlying model structure. On the other hand, if the range of model outputs encompasses observational data, it is not a guarantee that the model is conceptually correct.

All models are approximations, and it is only in use over time that we discover where the model diverges most from reality. Therefore, rather than think of models as something to accept or reject in an either/or fashion – to validate or invalidate – it may be more useful to think of models as tools to be modified in response to knowledge gained through continued observation of the natural systems being represented.

The inherent uncertainties in models of complex natural systems provide a strong argument for monitoring when models are used in support of public policy. Model predictions *will* be wrong. This is inescapable. We hope that they will be wrong in inconsequential ways, but we cannot guarantee this. Therefore, any action guided by model results – be it an artificial recharge system or a nuclear waste storage facility – should be subject to continued monitoring. The more serious the consequence of error, the more important such monitoring is. Modellers can play an important role in public policy as advocates for monitoring, by emphasising modelling as a

heuristic process, and by resisting the demand for predictions that are likely to be misleading, or simply wrong. Model output is not the same as a prediction of the future state of the system.

ACKNOWLEDGEMENTS

We are grateful to many colleagues for conversations that have helped to clarify the points made here, particularly William Ascher, John Bredehoeft, Dale Jamieson, Leonard Konikow, Daniel Sarewitz and Orrin Pilkey; and to an anonymous reviewer for thoughtful and helpful advice.

REFERENCES

Alley, W.M. and Emery, P.A. 1986. Groundwater model of the Blue River Basin, Nebraska – twenty years later. *Journal of Hydrology*, **85**, 225–249.

Anderson, M.P. and Woessner, W.W. 1992. The role of the postaudit in model validation. *Advances in Water Resources*, **15**, 167–173.

Anderson, T.W. 1968. Electric analog analysis of ground-water depletion in central Arizona. *US Geological Survey Water Supply Paper*, 1860, 21pp.

Aronica, G., Hankin, B. and Beven, K. 1998. Uncertainty and equifinality in calibrating distributed roughness coefficients in a flood propagation model with limited data. *Advances in Water Resources*, **22**, 349–365.

Ascher, W. 1978. *Forecasting: An Appraisal for Policy-Makers and Planners*. Johns Hopkins University Press, Baltimore, MD.

Ascher, W. 1981. The forecasting potential of complex models. *Policy Sciences*, **13**, 247–267.

Ascher, W. 1989. Beyond accuracy. *International Journal of Forecasting*, **5**, 469–484.

Ascher, W. 1993. The ambiguous nature of forecasts in project evaluation: diagnosing the over-optimism of rate-of-return analysis. *International Journal of Forecasting*, **9**, 109–115.

Ascher, W. and Overholt, W.H. 1983. *Strategic Planning and Forecasting: Political Risk and Economic Oppportunity*. Wiley, New York.

Beck, M.B., Ravetz, J.R., Mulkay, L.A. and Barnwell, T.O. 1997. On the problem of model validation for predictive exposure assessments. *Stochastic Hydrology and Hydraulics*, **11**, 229–254.

Bethke, C. 1992. The question of uniqueness in geochemical modelling. *Geochimica et Cosmochimica Acta*, **56**, 4315–4320.

Beven, K. 1993. Prophecy, reality, and uncertainty in distributed hydrological modelling, *Advances in Water Resources*, **16**, 41–51.

Beven, K. 1996. Equifinality and uncertainty in geomorphological modelling. In: Rhoads, B.L. and Thorn, C.E. (eds), *The Scientific Nature of Geomorphology: Proceedings of the 27th Binghampton Symposium in Geomorphology, 27–29 September 1996*. John Wiley, Chichester, 289–313.

Bredehoeft, J.D. and Konikow, L.F. 1992. Reply to comment (by de Marsily et al. 1992). *Advances in Water Resources*, **15**, 371–372.

Brunner, R.D. 1991. Global climate change: defining the policy problem, *Policy Sciences*, **24**, 291–311.

Brunner, R.D. and Ascher, W. 1992. Science and social responsibility. *Policy Sciences*, **25**, 295–331.

Dean, C. 1999. *Against the Tide: The Battle for America's Beaches*. Columbia University Press, New York.

de Marsily, G., Combes, P. and Goblet, P. 1992. Comment on 'Ground-water models cannot be validated', by L.F. Konikow and J.D. Bredehoeft. *Advances in Water Resources*, **15**, 367–369.

Fischhoff, B. 1982. For those condemned to study the past: heuristics and biases in hindsight. In: Kahneman, D., Slovic, P. and Tversky, A. (eds), *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge, 335–351.

Gupta, H.V., Sorooshian, S. and Yapo, P.O. 1998. Toward improved calibration of hydrologic models: multiple and noncommensurable measures of information. *Water Resources Research*, **34**, 751–763.

Gutierrez, G.J. and Kouvelis, P. 1991. Parkinson's Law and its implications for project management, *Management Science*, **37**, 990–1001.

Hall, M.J. and Pilkey, O.H. 1991. Effects of hard stabilization on dry beach width for New Jersey. *Journal of Coastal Research*, **7**, 771–785.

Hankin, B.G. and Beven, K.J. 1998a. Modelling dispersion in complex open channel flows: equifinality of model structure (1). *Stochastic Hydrology and Hydraulics*, **12**, 377–396.

Hankin, B.G. and Beven, K.J. 1998b. Modelling dispersion in complex open channel flows: fuzzy calibration (2). *Stochastic Hydrology and Hydraulics*, **12**, 397–412.

Hill, M.C. 1992. A computer program (MODFLOWP) for estimating parameters of a transient, three-dimensional, ground-water flow model using non-linear regression. *US Geological Survey Report*, 91–484.

Hillyer, T.M. and Stakhiv, E.Z. 1997. Discussion of Pilkey, O.H., 1996 (sic). The fox guarding the hen house (editorial). *Journal of Coastal Research*, **13**, 259–264.

Hodges, C.A. 1995. Mineral resources, environmental issues, and land use. *Science*, **268**, 1305–1312.

Houser, P.R., Shuttleworth, W.J., Famiglietti, J.S., Gupta, H.V., Syed, K.H. and Goodrich, D.C., 1998. Integration of soil moisture remote sensing and hydrologic modelling using data assimilation, *Water Resources Research*, **34**, 3405–3420.

Houston, J.R. 1990. Discussion of: Pilkey O.H., 1990. A time to look back at beach replenishment (editorial) and Leonard L., Clayton T. and Pilkey, O.H. 1990. An analysis of beach design parameters on US east coast barrier islands. *Journal of Coastal Research*, **6**, 1023–1036.

Houston, J.R. 1991. Beachfill performance. *Shore and Beach*, **59**, 15–24.

Isham, J. and Kaufmann, D. 1995. The forgotten rationale for policy reform: the productivity of investment projects. *The World Bank Policy Research Working Paper* 1549, 35pp.

Klemes, V. 1982. Empirical and causal models in hydrology. In: *National Research Council Geophysics Study Committee* (eds), *Scientific Basis of Water-Resource Management*. National Academy Press, Washington, DC, 95–104.

Konikow, L.F. 1986. Predictive accuracy of a ground-water model: lessons from a post-audit. *Ground Water*, **24**, 173–184.

Konikow, L.F. 1992. Discussion of 'The modelling process and model validation' by Chin-Fu Tsang. *Ground Water*, **30**, 622–623.

Konikow, L.F. and Bredehoeft, J.D. 1992. Ground-water models cannot be validated. *Advances in Water Resources*, **15**, 75–83.

Konikow, L.F. and Patten, E.P. Jr 1985. Groundwater forecasting. In: Anderson, M.G. and Burt, T.P. (eds), *Hydrological Forecasting*. John Wiley, Chichester, 221–270.

Konikow, L.F. and Person, M. 1985. Assessment of long-term salinity changes in an irrigated stream-aquifer system. *Water Resources Research*, **21**, 1611–1624.

Konikow, L.F. and Swain, L.A. 1990. Assessment of predictive accuracy of a model of artificial recharge effects in the upper Cochella Valley, California. In: Simpson, E.S. and Sharp, J.M. (eds), *Selected Papers on Hydrogeology from the 28th International Geological Congress (1989), volume 1*. Heinz Heise, Hannover, 433–449.

Leonard, L., Clayton, T. and Pilkey, O.H. 1990. An analysis of replenished beach design parameters on US east coast barrier islands. *Journal of Coastal Research*, **6**, 15–36.

Meadows, D.H., Meadows, D.L. and Randers, J. 1972. *The Limits to Growth: A Report for the Club of Rome's Project on the Predicament of Mankind*. Universe Books, New York.

Meadows, D.H., Meadows, D.L. and Randers, J. 1992. *Beyond the Limits: Confronting Global Collapse, Envisioning a Sustainable Future*. Chelsea Green Publishing Co, White River Junction, VT.

Moore, S. 1995. The coming age of abundance. In: Bailey, R. (ed.), *The True State of the Planet*. The Free Press, New York, 110–139.

Neumann, S.P. 1992. Validation of safety assessment models as a process of scientific and public confidence building. In: *High Level Radioactive Waste Management, Proceedings of the Third International Conference*. Las Vegas, NV, April 12–16, 1992, published by the American Nuclear Society, La Grange Park, Illinois, and the American Society of Nuclear Engineers, New York, 1404–1413.

Nigg, J. 2000. The issuance of earthquake 'predictions': scientific approaches and strategies. In: Sarewitz, D., Pielke, R. Jr and Byerly, R. Jr (eds), *Prediction: Science, Decision-Making and the Future of Nature*. Island Press, Washington, DC, 135–156.

Oreskes, N. 1998. Evaluation (not validation) of quantitative models. *Environmental Health Perspectives*,

106 (suppl. 6), 1453–1460.
Oreskes, N. 1999. *The Rejection of Continental Drift: Theory and Method in American Earth Science.* Oxford University Press, New York.
Oreskes, N., Shrader-Frechette, K. and Belitz, K. 1994. Verification, validation, and confirmation of numerical models in the earth sciences. *Science*, **263**, 641–646.
Peccei, A. 1977. *The Human Quality.* Pergamon Press, Oxford.
Pilkey, O.H. 1990. A time to look back at beach nourishment (editorial). *Journal of Coastal Research*, **6**, iii–vii.
Pilkey, O.H. Jr 1994. Mathematical modeling of beach behaviour doesn't work. *Journal of Geological Education*, **42**, 358–361.
Pilkey, O.H. 1995. The fox guarding the hen house. *Journal of Coastal Research*, **11**, iii–v.
Pilkey, O.H. 1997. Reply to: Hillyer, T.M. and Stakhiv, E.Z., 1997. Discussion of: Pilkey, O.H. 1996 (sic). The fox guarding the hen house (editorial). *Journal of Coastal Research*, **13**, 265–267.
Pilkey, O.H. 2000. Predicting the behavior of nourished beaches. In: Sarewitz, D., Pielke, R. Jr and Byerly, R. Jr (eds), *Prediction: Science, Decision-Making and the Future of Nature.* Island Press, Washington, DC, 159–184.
Pilkey, O.H. Jr and Thieler, E.R. 1992. Erosion of the United States Shoreline. *Quaternary Coasts of the United States: Marine and Lacustrine Systems, SEPM Special Publication*, **48**, 3–7.
Pilkey, O.H., Young, R.S., Riggs, S.R., Smith, A.W.S., Wu, H. and Pilkey, W.D. 1993. The concept of shoreface profile of equilibrium: a critical review. *Journal of Coastal Research*, **9**, 255–278.
Rayner, S. 2000. Prediction and its alternatives in climate change policy. In: Sarewitz, D., Pielke, R. Jr and Byerly, R. Jr (eds), *Prediction: Science, Decision-Making and the Future of Nature.* Island Press, Washington, DC, 269–296.
Sarewitz, D, Pielke, R.A., Jr. and Byerly, R., Jr. (eds) 2000. *Prediction: Science, Decision-making, and the Future of Nature.* Island Press, Washington, DC.
Sarewitz, D. and Pielke, R. Jr 1999. Prediction in science and society. *Technology in Society*, **21**, 121–133.
Simon, J.L. and Kahn, H. (eds.) 1984. *The Resourceful Earth: A Response to Global 2000.* Blackwell Scientific, Oxford.
Smith, A.W.S. 1990. Discussion of: Pilkey O.H., 1990. A time to look back at beach replenishment (editorial) and Leonard L., Clayton T. and Pilkey, O.H. 1990. An analysis of beach design parameters on US east coast barrier islands. *Journal of Coastal Research*, **6**, 1041–1045.
Steefel, C.I. and van Cappellen, P. 1998. Reactive transport modeling of natural systems. *Journal of Hydrology*, **209**, 1–7.
Swain, L.A. 1978. Predicted water-level and water-quality effects of artificial rechange in the upper Cochella Valley, California, using a finite-element digital model. *US Geological Survey Water Resources Investigation*, **77–29**, 54pp.
Tierney, J. 1992. Betting the planet. *The New York Times Magazine*, 2 December 1992, 52–81.
Tsang, C-F. 1991. The modeling process and model validation. *Ground Water*, **29**, 825–831.
Tsang, C-F. 1992. Reply to the preceding discussion of 'The modeling process and model validation'. *Ground Water*, **30**, 622–624.
Tversky, A. and Kahneman, D. 1982. Judgment under uncertainty: heuristics and biases. In: Kahneman, D., Slovic, P. and Tversky, A. (eds), *Judgment under Uncertainty: Heuristics and Biases.* Cambridge University Press, Cambridge, 3–20.
Young, R.S., Pilkey, O.H., Bush, D.M. and Thieler, E.R. 1995. A discussion of the Generalized Model for Simulating Shoreline Change (GENESIS). *Journal of Coastal Research*, **11**, 875–886.