2

The Role of Quantitative Models in Science

Naomi Oreskes

Summary

Models in science may be used for various purposes: organizing data, synthesizing information, and making predictions. However, the value of model predictions is undermined by their uncertainty, which arises primarily from the fact that our models of complex natural systems are always open. Models can never fully specify the systems that they describe, and therefore their predictions are always subject to uncertainties that we cannot fully specify. Moreover, the attempt to make models capture the complexities of natural systems leads to a paradox: the more we strive for realism by incorporating as many as possible of the different processes and parameters that we believe to be operating in the system, the more difficult it is for us to know if our tests of the model are meaningful. A complex model may be more realistic, yet, ironically, as we add more factors to a model, the certainty of its predictions may decrease even as our intuitive faith in the model increases. For this and other reasons, model output should not be viewed as an accurate prediction of the future state of the system. Short timeframe model output can and should be used to evaluate models and suggest avenues for future study. Model output can also generate "what if" scenarios that can help to evaluate alternative courses of action (or inaction), including worst-case and best-case outcomes. But scientists should eschew long-range deterministic predictions, which are likely to be erroneous and may damage the credibility of the communities that generate them.

The Role of Quantitative Models in Science

What is the purpose of models in science? This general question underlies the specific theme of this volume: What should be the role of quantitative models in ecosystem science? Ultimately the purpose of modeling in science must be congruent with the purpose of science itself: to gain understanding of the natural world. This means understanding both processes and products, the things in the world and the ways in which they interact. Historically, scientists have sought understanding for many reasons: to advance utilization of earth resources, foster industrialization, improve instruments and techniques of warfare, prevent or treat disease, generate origins stories, reflect on the glory and beneficence of the world's creator, and satisfy human curiosity. None of these goals has proved itself superior to any of the others; science has advanced under all of these motivations.

Until the twentieth century, the word "model" in science typically referred to a physical model—such as the seventeenth-century orreries built to illustrate the motions of the planets. Physical models made abstract ideas concrete and rendered complex systems visualizable, enabling scientists to think clearly and creatively about complex systems. Models provided analogies: in the early twentieth century, the orbital structure of the solar system provided a cogent analogy for the orbital structure of the atom. Models also served as heuristic devices, such as the nineteenth-century mechanical models that used springs and coils to interrogate the mechanism of light transmission through the ether, or the twentieth-century wooden models of sticks and balls used to explore possible arrangements of atoms within crystal structures. Perhaps most ambitiously, physical models were used as arguments for the plausibility of proposed causal agents, such as the nineteenth-century scale models of mountains, in which various forms of compression were applied in hopes of simulating the fold structures of mountain belts to demonstrate their origins in lateral compression of the earth.

In our time, the word "model" has come to refer primarily to a computer model, typically a numerical simulation of a highly parameterized complex system. The editors of this volume suggest that quantitative models in ecosystem science have three main functions: synthesis and integration of data; guiding observation and experiment; and predicting or forecasting the future. They suggest that scientists are well aware of the value of models in integrating data and generating predictions, but are less well informed about the heuristic value of models in guiding observation and experiment. In their words,

> Quantitative models provide a means to test our understanding of ecosystems by allowing us to explore the interactions among observations, synthesis, and prediction. The utility of models for synthesis and prediction is obvious. The role of quantitative models in informing and guiding observation and experimentation is perhaps less often appreciated, but equally valuable. (Canham et al. 2001)

There is, however, a generous literature on the heuristic value of models and their use in guiding observation and experiments, particularly in the physical

sciences, which need not be reiterated here (e.g., Cartwright 1983; Tsang 1991, 1992; Konikow 1992; Konikow and Bredehoeft 1992; Beven 1993, 2000, 2001, 1999; Oreskes et al. 1994, Rastetter 1996; Narisimhan 1998; Oreskes 1998; Morgan and Morrison 1999). The focus of this essay is therefore on challenging the "obvious"—that is, challenging the utility of models for prediction.

To be sure, many scientists have built models that can be run forward in time, generating model output that describes the future state of the model system. Quantitative model output has also been put forward as a basis for decision making in socially contested issues such as global climate change and radioactive waste disposal. But it is open to question whether such models generate reliable information about the future, and therefore in what sense they could reasonably inform public policy (Oreskes et al. 1994; Pilkey 1994; Shackley et al. 1998; Evans 1999; Morgan 1999; Sarewitz and Pielke 2000, Sarewitz et al. 2000; Oreskes 2000a; Oreskes and Belitz 2001). Moreover, it is not even clear that time-forward model output necessarily contributes to basic scientific understanding. If our goal is to understand the natural world, then using models to predict the future does not necessarily aid that goal. If our goal is to contribute usefully to society, using models to predict the future may not do that, either.

Why should we think that the role of models in prediction is obvious? Simply because people do something does not make its value obvious; humans do many worthless and even damaging things. To answer the question of the utility of models for prediction, it may help to step back and think about the role of prediction in science in general. When we do so, we find that our conventional understanding of prediction in science doesn't work for quantitative models of complex natural systems precisely because they are complex. The very factors that lead us to modeling—the desire to integrate and synthesize large amounts of data in order to understand the interplay of various influences in a system—mitigate against accurate quantitative prediction.

Moreover, successful prediction in science is much less common than most of us think. It has generally been limited to short-duration, repetitive systems, characterized by small numbers of measurable variables. Even then, success has typically been achieved only after adjustments were made based on earlier failed predictions. Predictive success in science, as in other areas of life, usually ends up being a matter of learning from past mistakes.

Models Are Open Systems

The conventional understanding of scientific prediction is based on the hypothetico-deductive model. Philosophers call it the deductive-nomological model, to convey the idea that, for our predictions to be correct, they must derive from stable scientific laws. Whatever one prefers to call it, this model assumes that the principal task of science is to generate hypotheses, theories, or laws and compare their logical consequences with experience and observations in the natural world. If our predictions match the facts of the world, then we can say that the hypothesis has been confirmed, and we can feel good about what we have done so far. We are on the right track. If the observations and experiences don't match our predictions, then we say that the hypothesis has been refuted and we need to go back and make some adjustments.

The problem with the hypothetico-deductive model, as many scientists and philosophers have realized, is that it works reliably only if we are dealing with closed systems. The hypothetico-deductive model is a logical structure of the form "$p \rightarrow q$," where our proposition, p, does in fact entail q, if and only if $p \rightarrow q$ is a complete description of the system. That is, if and only if the system is closed. But natural systems are never closed: they always involve externalities and contingencies that may not be fully specified, or even fully known. When we attempt to test a hypothesis in the real world, we must invoke auxiliary assumptions about these other factors. This means all the additional assumptions that have to be made to make a theory work in the world: frictionless surfaces, ideal predators, purely rational humans operating in an unfettered free market. When we test a theory by its consequences, other potentially influential factors have to be held constant or assumed not to matter. This is why controlled experiments play such a pivotal role in the scientific imagination: in the laboratory we have the means to control external conditions in ways that are not available in ordinary life. Yet even in the laboratory, we still must assume—or assert—that our controlled factors are in fact fully controlled and that the factors we consider negligible are in fact so. If a theory fails its test, we cannot be certain whether the fault lies in the theory itself or in one of our other assumptions.

Another way to understand this is to compare a model of a natural system with an artificial system. For example, in our commonly used arithmetic, we can be confident that if $2 + 2 = 4$, then $4 - 2 = 2$, because we have *defined* the terms this way and because no other factors are relevant to the problem. But consider the question: Is a straight line the shortest distance between two points? Most of us would say yes, but in doing so we would have invoked the auxiliary assumption that we are referring to a planar surface. In the abstract world of Euclidian geometry, or on a high-school math test, that would be a reasonable assumption. In high school, we'd probably be classified as a smart aleck if we wrote a long treatise on alternative geometrical systems. But if we are referring to natural systems, then we need additional information. The system, as specified, is open, and therefore our confident assertion may be wrong. The shortest distance between two points can be a great circle.

Furthermore, in order to make observations in the natural world, we invariably use some form of equipment and instrumentation. Over the course of history, the kinds of equipment and instruments that have been used in science have tended to become progressively more sophisticated and complex. This means that our tests have become progressively more complex, and apparent failures of theory may well be failures of equipment, or failures on our part to understand the limitations of our equipment.

The most famous example of this in the history of science is the problem of stellar parallax in the establishment of the heliocentric model of planetary motions (Kuhn 1957). When Nicolaus Copernicus proposed that model in the early sixteenth century, it was widely recognized that this idea would have an observable consequence: stellar parallax—the apparent changing position of a star in the heavens as the earth moved through its orbit. If the earth stood still, its position relative to the stars would be constant, but if it moved, then the position of the stars would seem to change from winter to summer and back again. One could test the Copernican model by searching for stellar parallax. This test was performed and no parallax was found, so many astronomers rejected the theory. It had failed its experimental test.

Four hundred years later, we can look back and see the obvious flaw in this test: it involved a faulty auxiliary hypothesis—namely, that the universe is small. The test assumes that the diameter of the earth's orbit around the sun is large relative to the distance to the star and that the stellar parallax is a large angle. Today, we would say this is a conceptual flaw: the stars are almost infinitely far away and the parallax is therefore negligibly small.

The experimental test of stellar parallax also involved an assumption about equipment and instrumentation: namely, that the available telescopes were accurate enough to perform the test. Today, astronomers can detect stellar parallax, which is measurable with the instruments of the twenty-first century, but sixteenth-century telescopes were simply inadequate to detect the change. Of course, our equipment and instrumentation are far more sophisticated today, but the same kinds of assumptions of instrumental adequacy are built into our tests as were built into theirs.

This brings us to the kind of models that most of us work with today. The word "model" can be problematic because it is used to refer to a number of different things, but this discussion will assume that we are referring to a mathematical model, typically a numerical simulation, realized on a digital computer. However, the points may apply to other kinds of models as well.

All models are open systems. That is to say, their conclusions are not true by virtue of the definition of our terms, like "2" and "+," but only insofar as they encompass the systems that they represent. Alas, no model completely encompasses any natural system. By definition, a model is a simplification—an idealization—of the natural world. We simplify problems to make them tractable, and the same process of idealization that makes problems tractable also makes our models of them open. This point requires elaboration.

There are many different ways in which models are open, but there are at least three general categories into which this openness falls. First, our models are open with respect to their conceptualization (how we frame the problem). When we create a model, we abstract from the natural world certain elements that we believe to be salient to the problem we wish to understand, and we omit everything else. Indeed, good science *requires* us to omit irrelevancies. For example, consider a model of predator-prey relations in the Serengeti Plain. We can be fairly confident that the color of my bedroom is irrelevant to this model.

There is no known reason why it should matter to animals in Africa; indeed, there is no way (so far as we know) that the animals could even be aware of it. But a little imagination reveals that there could be other factors that we consider irrelevant but that might in the future be shown to matter. (The history of science is full of connections and correlations that were previously unsuspected but later demonstrated. It is also full of examples of correlations that were later rejected.) Moreover, there may be factors that we know or suspect *do* matter, but which we leave out for various reasons—we lack time, computational power, or other resources to incorporate them; we lack data or analytical methods to represent them; or we lack confidence about their significance (Ascher 1993; Oreskes and Belitz 2001). Or we may simply frame the problem incorrectly (for a more detailed discussion, with specific examples of this, see Oreskes and Belitz 2001). At every level there remains the question whether the model conceptualization is adequate.

Our models are also open with respect to the empirical adequacy of the governing equations (how well our mathematical representations map onto natural processes). We often call these equations *laws*, but as philosopher Nancy Cartwright (1983) has cogently shown, this usage is really rather misleading. Scientific laws are idealizations that map onto the natural world to a greater or lesser degree depending on the circumstances. Moreover, while the use of the term "law" was borrowed from the political realm, laws of nature are different from laws of the state.

Political laws do not attempt to describe an actual state of affairs, but rather the opposite: they announce how we want things to be. Political laws are declarations of intent, and we adjudicate them based on rather formal procedures. Laws of nature are not declarations. They are our best approximations of what we think is really going on, and there is no formal standard by which we judge them. In a numerical model, we assume, based on prior experience, that the equations used are adequate, but we have no logical way to demonstrate that this assumption is correct. Therefore the model remains open.

Third, models are open with the respect to the input parameterization (how well numerical variables represent elements of the system). In some systems, like planets orbiting around the sun, we can count the number of objects involved in the problem—nine planets[1] and the sun—and make our best measurements of their size, shape, distance, etc. But in many other problems, it is an open question as to exactly what the relevant variables are. Even if we know what they are—for example, permeability in a hydrological model—we still have to make choices about how best to represent and parameterize it.

It is important to underscore that calling a model open is not the same as calling it "bad." The more we don't know about a system, the more open our model will be, even if the parts we do know about are very well constrained. A card game is often taken as a good example of a closed system because we know how many cards are in the deck and precisely what those cards have to be, assuming no one is cheating. But there's the rub. To be confident of our odds, we have to make the assumption of honest dealing. In fact, this paradigmatically closed system is, in real life, open. Poker is a good game, but in real life it's an open system. Similarly, it is possible to imagine a model in which the available empirical data are well measured and in which the governing equations have stood the test of time, yet in which important relevant parameters have not yet been recognized. Such a model might be "good" in terms of the standards of current scientific practice yet remain highly open and therefore fail to make reliable predictions.

The Complexity Paradox

The openness of a model is a function of the relationship between the complexity of the system being modeled and the model itself. The more complex the natural system is, the more different components the model will need to mimic that system. Therefore we might think that by adding components we can make the model less open. But for every parameter we add to a model, we can raise a set of questions about it: How well does it represent the object or process it purports to map onto? How well constrained is our parameterization of that feature? How accurate are our measurements of its specific values? How well have we characterized its interrelations with other parameters in the model? Even as we increase our specifications, the model still remains an open system.

This might suggest that simpler models are better—and in some cases no doubt they are—but in ecosystems modeling we do not want to abandon complexity because we believe that the systems we are modeling are in fact complex. If we can demonstrate that certain parameters in the model are insignificant, then we can omit them, but in most cases that would be assuming the thing we wish to discover: What role does this parameter play? How does it interact with other parameters in the system? Indeed, in many cases it is the very complexity of the systems that has inspired us to model them in the first place—to try to understand the ways in which the numerous parts of the system interact.

Moreover, complexity can improve accuracy by minimizing the impact of errors in any one variable. In an analysis of ecological models of radionuclide kinetics in ecosystems, O'Neill (1973) showed, as one might expect, that systematic bias resulting from individual variables decreased as the number of variables in a model increased. However, uncertainty *increased* as the measurement errors on individual parameters accumulated. Each added variable added uncertainty to the model, which, when promulgated in a Monte Carlo simulation, contributed to the uncertainty of the model prediction.

These considerations may be summarized as the "complexity paradox." The more we strive for realism by incorporating as many as possible of the different processes and parameters that we believe to be operating in the system, the more difficult it is for us to know if our tests of the model are meaningful. Put another way, the closer a model comes to capturing the full range of processes and parameters in the system being modeled, the more difficult it is to ascertain whether or not the model faithfully represents that system. A complex model may be more realistic yet at the same time more uncertain. This leads to the ironic situation that as we add more factors to a model, the certainty of its predictions may decrease even as our intuitive faith in the model increases. Because of the complexities inherent in natural systems, it may never be possible to say that a given model configuration is factually correct and, therefore, that its predictions will come true. In short, the "truer" the model, the more difficult it is to show that it is "true."

Successful Prediction in Science

At this point some readers will be thinking, "But surely there are many cases in which scientists have made successful predictions, many areas of science where we do a good job and our predictions have come true." This intuition may be recast as a question: Where *have* scientists developed a track record of successful prediction? What do we learn when we examine the nature of those predictions and how they have fared? Viewed this way, we find a surprising result: successful prediction in science is less common than most of us think, and it has developed as much through a process of trial and error as through the rigorous application of scientific law. Consider three areas of science that have a large literature on their predictive activities: weather, astronomy, and classical mechanics.

Example 1: Meteorology and Weather Prediction

Meteorology is the science most closely associated with prediction in the public mind and the only science that regularly earns time on the evening news. What do we know about weather prediction? First, that it is non-deterministic. Weather forecasts are not presented in the form, "It will rain two inches tomorrow beginning at 3:00 o'clock and lasting until 4:30." They are presented in the form, "There is a 20% chance of rain tomorrow afternoon to early evening." Moreover, if rain is expected, forecasters typically offer a range of plausible values, such as 1–2 inches. In a sense, we could say that meteorologists hedge their bets, and there has been considerable debate within the meteorological community over just how weather forecasts should be presented (e.g., Murphy 1978). The use of probabilistic forecasting is partly a response to experience: history has demonstrated just how difficult specific, quantitative prediction of a complex system is.

Second, weather prediction involves spatial ambiguity. If my local forecast calls for an 80% chance of rain tomorrow in San Diego County, that forecast will be deemed accurate if it does in fact rain, even though some parts of the county may remain dry. Some of us have seen the phenomenon where it is raining at our house and dry across the street. Weather can be very local; forecasts are spatially averaged; and typically they become more accurate as they become more general and less accurate as they attempt greater specificity.

Third, and perhaps most important, accurate weather prediction is restricted to the near term. In meteorology, a "long-range" forecast is 3–5 days (or lately perhaps a bit longer.) The great successes in the history of meteorology, like the celebrated D-Day forecasts, are a case in point (Petterssen 2001). If you need an accurate forecast for the weather on January 27 next year, you simply cannot get it. (Anyone who has tried to plan an outdoor wedding is familiar with this problem. In fact there are now folks who *will* sell you a forecast for next June 10, and there are also folks who still sell snake oil.)

Partly in response to this problem, atmospheric scientists have developed a distinct terminology to deal with long-term change—they speak of (general) climate rather than (specific) weather. Meteorologists can accurately predict the average temperature for the month of January because the dominant control on monthly weather is the annual journey of Earth around the Sun, coupled with the tilt of Earth's axes, factors that are themselves quite predictable. Yet these planetary motions are not the only relevant variables: natural climate variation also depends on solar output, gases and dust from volcanoes, ocean circulation, perhaps even weathering rates controlled by tectonic activity. These latter processes are less regular than the planetary motions and therefore less predictable. Hence our predictive capacity is constrained to the relatively near future: we can confidently predict the likely average temperature of for a specific month within the next couple of years, but not so the average temperature of that month in 2153 (claims to the contrary not withstanding). The more extended the period, the more difficult the forecasting task becomes.

We know something about why long-range weather forecasting is so difficult: weather patterns depend upon external forcing functions, such as the input of solar radiation, small fluctuations in which can produce large fluctuations in the system. Weather systems are also famously chaotic, being highly sensitive to initial conditions. Many systems of interest to ecologists are similar to climate: they are strongly affected by exogenous variables. We know that these factors are at play, and we may even understand the reasons why they vary, but we cannot predict how their variations will alter our systems in the years to come. This is why a model can be useful to guide observation and experiment yet be unable to make accurate predictions: we can use a model to test which of several factors is most powerful in affecting the state of a system and use that information to motivate its further study, but we cannot predict which of these various factors will actually change and therefore what the actual condition of the system will be.

There is an additional point to be made about weather prediction. The reason we can make accurate predictions at all is because our models are highly calibrated. They are based on enormous amounts of data collected over an extended time period. In the United States and Europe, weather records go back more than a century, and high quality standardized records exist for at least four decades. In the United States, there are 10 million weather forecasts produced by the National Weather Service each year (Hooke and Pielke 2000)! Predictive weather models have been repeatedly adjusted in response to previous bad predictions and previous failures. Compare this situation with the kinds of models currently being built in aid of public policy: general circulation models, advective transport models, forest growth models, etc. None of these has been subject to the kind of trial and error—the kind of learning from mistakes—that our weather models have been.

We can readily learn from mistakes in weather modeling because weather happens every day and because of the enormous, worldwide infrastructure that has been created to observe and forecast weather. A similar argument can be made about another area of successful prediction in science: celestial mechanics.

Example 2: Celestial Mechanics and the Prediction of Planetary Motions

Celestial mechanics is an area in which scientists make very successful predictions. The motion of the planets, the timing of comets and eclipses, the position and dates of occultations—scientists routinely predict these events with a high degree of accuracy and precision. Unlike rain, we can predict to the minute when a solar eclipse will occur and precisely where the path of totality will be. Given this success, we can ask ourselves: What are the characteristics of these systems in which we have been able to make such successful predictions?

The answer is, first, that they involve a small number of measurable parameters. In the case of a solar eclipse, we have three bodies—Earth, its moon, and the Sun—and we need to know their diameters and orbital positions. This is a relatively small number of parameters, and each has a fixed value. Put another way, the variables in the system do not actually vary. The diameter of Earth is not, for all intents and purposes, changing (or at least not over the time frame relevant to this prediction). Second, the systems involved are highly repetitive. Although eclipses don't happen every day, they happen a lot, and we can track the positions of the planets on a daily basis. Moreover, planets return to their orbital positions at regular intervals. When we make an astronomical prediction, we can compare it with observations in the natural world, generally without waiting too long. If our predictions fail because we have made a mistake, we can find this out fairly quickly and make appropriate adjustments to the model. Third, as in the case of weather, humans have been making and recording observations of planetary motions for millennia. We have an enormous database with which to work.

However, precisely because of its successful track record, the development of accurate planetary prediction raises one of the most serious concerns for modelers: that faulty models may make accurate predictions. Returning to Copernicus, the failure of the prediction of stellar parallax was a serious problem for the heliocentric model, but there was another problem more pressing still: the Ptolemaic system, which the heliocentric model aspired to replace, made highly accurate predictions. In fact, it made *more* accurate predictions than the Copernican system until many decades later when Johannes Kepler introduced the innovation of elliptical orbits. Astronomers did not reject the Copernican view because they were stubborn or blinded by religion. They rejected it because the model they were already working with was better, as judged by the accuracy of its predictions. Yet in retrospect we believe that it was wrong at its very foundations.

This should give any modeler pause, for it shows that accurate predictions are not proof of a model's conceptual veracity: false models can make true predictions. In this case, repeated accurate predictions serve to bolster what in hindsight will be viewed as misplaced confidence. Conversely, a model that is conceptually well grounded may still fail in its predictions if key elements of the model are incorrect. Therefore, building a model that makes accurate predictions may not necessarily aid the goal of improved basic understanding. Indeed, it may actually impede it.

Many modelers achieve successful prediction through a process of model calibration: they obtain observations from the natural world and adjust model parameters until the model successfully reproduces the observational data. But this kind of calibration ensures that the model cannot fail—effectively, it makes the model refutation-proof (Oreskes et al. 1994; see also Anderson and Bates 2001). If there is a conceptual flaw in the model, unconstrained calibration will hide the flaw from view. This is what the Ptolemaic astronomers did: they added epicycles upon epicycles to "save the phenomenon"—to make the model fit the data—all the while preserving a system that was conceptually flawed at its root.

Example 3: Classical Mechanics

A third example may be drawn from classical mechanics. This is the area of science that most closely matches the hypothetico-deductive ideal of deterministic laws that generate specific quantitative predictions. We all learned these laws in high school physics: $f = ma$, $s = \frac{1}{2} at^2$, $\rho = mv$, etc. What do we really know about these laws, which most of us think of as the best example of deterministic laws that make accurate predictions? None is literally true. To generate accurate predictions, they all require *ceteris paribus* clauses: that there is no friction, no air resistance, or that an ideal ball is rolling down a perfectly frictionless plane. In real life, the laws do not work as stated, because in real life these kinds of ideal systems do not exist. Philosopher Nancy Cartwright calls this how the laws of physics "lie"—they posit an imaginary world, not the world that we live in (Cartwright 1983; see also Cartwright 1999). To make the laws of physics work in practice requires adjustments and modifications that are *not* based upon deterministic laws but rather on past experience and earlier failed attempts. So what starts out looking like a successful case of clean, deterministic prediction becomes a great deal messier as you get down into the nitty-gritty details.

Where does this leave us? First, we can draw the conclusion that accurate prediction in science is a special case (Cartwright 1999). It tends to be restricted to systems characterized by small numbers of measurable parameters or to systems in which the events at issue are naturally repetitive, like the orbits of the planets, or easily repeated, like balls rolling down hills. Second, accurate predictions have usually been achieved only after adjustments to the model in response to earlier failed predictions. This is why the interplay between modeling and observation is so important. It also is why repetitive systems are more likely to be predictable even if they are chaotic. Whether a system is intrinsically deterministic or intrinsically chaotic, we are more likely to be able to predict its behavior successfully if we have observed it many times before. Third, even when we achieve a match between quantitative prediction and empirical observation, it does not guarantee an accurate conceptualization of a system. Faulty conceptual models can make accurate predictions. Predictive capacity by itself is a weak basis for confidence.

Model Testing, Forecasting, and Scenario Development

Our focus here has been on branches of science that have a track record of prediction—meteorology, astronomy, and physics. The natural sciences that deal with complex nonrepetitive systems—geology, biology, and ecology—have no historic track record of predictive success at all. Indeed, until very recently, no one ever expected scientists in these disciplines to make predictions (Oreskes 2000b). If our science has no track record of predictive success, then it behooves us to ask, what makes us think we are different?

To answer this question, it may help to distinguish between short-term and long-term prediction, or what might be better referred to as model testing versus forecasting. Clearly, short-term predictions that can be used as a means to evaluate the model are extremely important, because they can provide a test of our understanding. The Copernican example shows that agreement between prediction and observation is not proof that a model is right, but disagreement can be evidence that something is wrong. Moreover, short-term predictions can be used to compare alternative models: to give us a handle on which of a number of possible conceptualizations or parameterizations does a better job of describing a given system, or to see what range of outcomes is suggested by current understandings. The latter in particular may be highly relevant for policy decisions, for example, in developing worst-case scenarios of climate change. Understood this way, prediction—that is to say, short-term prediction—becomes one of the heuristic functions of modeling.

However, long-range predictions—or *forecasts*—cannot be tested, and therefore do nothing to improve scientific understanding. A concrete example may help to clarify this point. Recently, the U.S. Department of Energy has issued forecasts of radionuclide releases at the proposed nuclear waste repository at

Yucca Mountain ten thousand years from now. These projections are interpreted to be required by law to ensure that the repository will satisfy regulatory standards. If the law demands such forecasts, then they will be made, and they are not necessarily useless: if the values are considered too high they may inspire modification to the repository design, or perhaps even stop the repository from going forward. In this way, such forecasts can be socially and politically useful (or harmful, depending upon your point of view). But from a scientific point of view, such long-range forecasts offer very little value, for they cannot be compared with empirical results and therefore cannot be used to improve scientific knowledge. If we understand our purpose as fundamentally a scientific one—to improve our understanding of natural systems—then long-range forecasting cannot aid this goal. Only when predictions are on a short enough time frame that we can compare them to events in the natural world can they play a role in improving our comprehension of nature.

This brings us to the domain of models and public policy. The desideratum of policy-relevant scientific information is frequently cited as a major motivation for modeling in ecology, atmospheric science, geochemistry, hydrology, and many other areas of natural and social science (Oreskes et al. 1994, Oreskes 1998). Several papers in this volume state that one of the important reasons for building the model under discussion is its relevance to public policy.

Some questions of public policy involve short time frames, and these questions may be usefully informed by model results. However, questions about short-duration events are typically questions about particular localities whose details are unique (Beven 2000). Given the uncertainties of the details of most site-specific models, predictive modeling is unlikely to be a substitute for monitoring. The problem here, as recently emphasized by Anderson and Bates (2001), is that data collection and monitoring are frequently viewed as "mundane," whereas modeling is viewed as "cutting-edge." Perhaps for this reason, data-collection programs have proved difficult to sustain. Scientists should ponder why this is so and consider whether the public interest is being served by our emphasis on models. A better balance between modeling and data collection may be called for.

Better data collection for model testing and monitoring may improve the usefulness of models used for short-term policy decisions, but the policy issues that models are commonly held to illuminate often involve long-term change in natural systems. Global climate change is the most obvious, but many, perhaps most, ecological models involve questions whose import will be realized over time frames of at least years, if not decades. Yet all of the available evidence suggests that long-term forecasts are likely to be wrong and may very well misinform public policy.

Given how many models have been built in the past decades, it is remarkable how few have been evaluated after the fact to determine whether their forecasts came true. Where such post hoc evaluations have been done, the results are extremely disconcerting: the vast majority of the predictions failed, often they failed rapidly, and in many cases the actual results were not only outside the error bars of the model, but they were in an entirely different direction (Ascher 1978, 1981, 1989, 1993; Ascher and Overholt 1983; Konikow 1986; Konikow and Patten 1985; Konikow and Person 1985; Konikow and Swain 1990; Leonard et al. 1990; Pilkey 1990, 1994, 2000; Nigg 2000; see also Sarewitz et al. 2000 and Oreskes and Belitz 2001).

The most detailed work on the question of forecasting accuracy has been done by political scientist William Ascher, working on economic and social science models (Ascher 1978, 1981, 1989, 1993; Ascher and Overholt 1983). For example, in a study of the rate of return on development projects, he found that in most cases the return was not only lower than predicted, but it was below the designated cut-off criterion for the funding agency. Indeed, it was so low that had it been accurately predicted, the project would never have been funded in the first place (Asher 1993). Another example is the celebrated "World Model" of the Club of Rome, published in the 1970s best-seller, *The Limits to Growth* (Meadows et al. 1972). This model predicted major shortages of natural resources by the end of the twentieth century that would cause commodity prices to skyrocket. In fact, not only did prices fail to increase at the predicted rate, they did not increase at all. At the end of the twentieth century, the prices for nearly all natural resources were lower than at the time the model was built (Moore 1995).

Natural scientists may be inclined to discount results from the social sciences because we have been trained to think of the systems they deal with as poorly constrained and subject to the intrinsic unpredictability of human behavior. Yet if one considers why social sciences models are messy—particularly the difficulty of specifying and quantifying system variables—we find much in common with models of complex natural systems. Moreover, if part of the motivation for our model-building is policy-relevance, then by definition we are looking at systems which in some way either affect or have been affected by humans, in which case the unpredictability of human behavior may well be relevant. For example, in his post audit of hydrological models, Leonard Konikow found that models typically failed because of unanticipated changes in the forcing functions of the system, and these forcing functions commonly involved human activities such as groundwater pumping, irrigation, and urbanization (Konikow 1986).

Faulty Forecasts Undermine Scientific Credibility

If the forecasts of a model are wrong, then sooner or later they *will* be refuted. For complex systems that are poorly constrained, or for which there are limited historical data, experience suggests that sooner is the more likely timeframe. If the forecasts are presented as facts—and they typically are—then this ultimately undermines the credibility of the community that generated them. A good example of this comes from the 1997 Grand Forks flood, an example of prediction gone wrong that has been analyzed by Pielke (1999) and Changnon (2000).

In the spring of 1997, the Red River of the North, which flows past Grand Forks, North Dakota, crested at the historically unprecedented level of 54.11 feet, causing massive flooding and $1–2 billion in property damage. Several weeks before, the National Weather Service had issued two flood "outlooks," one based on a scenario of average temperature and no additional precipitation, the other based on average temperature with additional precipitation. The two outlooks were 47.5 and 49 feet, respectively. Because spring flooding in this region is largely controlled by snowmelt, one might expect such forecasts to be reliable—after all, the snowpack can be measured. So the town of Grand Forks prepared. Residents added a couple of feet "just in case" and prepared for flood crests of up to 51 feet (Pielke personal comm. 2002). When the waters rose to 54 feet, massive damage ensued.

Although the scientists involved surely understood the uncertainties in their forecasts, and the Weather Service appended qualitative disclaimers, these uncertainties were not "received" by local officials. Rather they interpreted the outlooks as "facts"—either that the river would crest between 47.5 and 49 feet or that it would crest no higher than 49 feet. As these numbers were repeated by local officials, the media, and ordinary citizens, their status as facts hardened. When these "facts" proved false, blame for the disaster was laid at the feet of the Weather Service. While one might argue that local officials were at least in part responsible—for misunderstanding the scientific data and not preparing for worst-case scenarios—or that this was simply a tragic but uncontrollable natural disaster—an act of God or other unforeseeable event—this is not how the disaster was interpreted by those involved. Rather, it was interpreted as a failure on the part of the Weather Service. As Pielke (1999) quotes the mayor of Grand Forks, "they blew it big."

Examples like these illustrate how scientists can and will be blamed for faulty predictions, even if disclaimers or error bars are included and even if scientists have done their best in the face of an unprecedented natural event. Given this, it is hard to see how it can be in the long-term interest of the scientific community to make confident assertions that will be soon refuted. At best, the result is embarrassment; at worst, the risk is a loss of confidence in the significance and meaning of scientific information and ultimately a loss of public support for our endeavors.

The alternative is to shift our focus away from specific quantitative predictions of the future and toward policy-relevant statements of scientific understanding (see Sarewitz and Pielke 2000, and this volume chapter 7, for related views). Rather than attempt a specific prediction that is likely to fail, we can develop "what if" scenarios that highlight, for policy-makers and other interested parties, what the most likely consequences will be of alternative possible courses of action. To say that something is policy-relevant is to say that there is some possible course of human action (or inaction) that bears on the future state of affairs and that our information might plausibly affect which course of action humans will take.

We cannot say what will happen in the future, but we can give an informed appraisal of the possible outcomes of our choices. In the case of Grand Forks, Pielke argues that the National Weather Service intended to convey—and thought it had conveyed—the message that the communities involved should prepare for unprecedented flooding. Imagine then, for a moment, if instead of issuing their quantitative outlooks, the Weather Service had said just this: "All the available evidence suggests that we need to prepare for unprecedented flooding." The outcome, both for the reputation of the Weather Service and, more importantly, for the people of Grand Forks, might have been very different.

Acknowledgments. I am grateful to the organizers of Cary Conference IX for inviting me to present my ideas there and in this volume, and to Roger Pielke Jr. and an anonymous reviewer for very helpful reviews. This paper represents a summary and amalgamation of ideas presented and published elsewhere over the past several years; I am indebted to Kenneth Belitz, John Bredehoeft, Nancy Cartwright, Leonard Konikow, Dale Jamieson, Roger Pielke Jr., and Daniel Sarewitz for ongoing conversations about the meanings and purposes of modeling, and to Dennis Bird, who first inspired me to think hard about experience and observation in the natural world. I completed the revisions to this paper while visiting in the Department of History of Science, Harvard University, where I was generously provided with an ideal environment for thinking and writing about science in all its dimensions.

Note

1. Or maybe eight, depending upon your opinion of the status of Pluto.

References

Anderson, M.G., and P.D. Bates. 2001. Hydrological science: Model credibility and scientific integrity. Pp. 1–10 in M.G. Anderson and P.D. Bates, editors. *Model Validation: Perspectives in Hydrological Science.* Chichester: John Wiley and Sons.

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

———. 1981. The forecasting potential of complex models. *Policy Sciences* 13: 247–267.

———. 1989. Beyond accuracy. *International Journal of Forecasting* 5: 469–484.

———. 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 W.H. Overholt. 1983. *Strategic Planning and Forecasting: Political Risk and Economic Opportunity.* New York: John Wiley and Sons.

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

———. 2000. Uniqueness of place and process representations in hydrological modeling. *Hydrology and Earth Systems Science* 4: 203–213.

———. 2001. Calibration, validation and equifinality in hydrological modeling: A continuing discussion. Pp. 43–55 in M.G. Anderson and P.D. Bates, editors. *Model Validation: Perspectives in Hydrological Science.* Chichester: John Wiley and Sons.

Canham, C.D., J.J. Cole, and W.K Lauenroth. 2001. Understanding ecosystems: The role of quantitative models in observation, synthesis, and prediction. Paper presented at Cary Conference IX, May 1–3, 2001, Institute of Ecosystems Studies, Millbrook, NY.

Cartwright, N. 1983. *How the Laws of Physics Lie.* Oxford: Clarendon Press.

———. 1999. *The Dappled World: A Study of the Boundaries of Science.* Cambridge: Cambridge University Press.

Changnon, S.A. 2000. Flood prediction: Immersed in the quagmire of national flood mitigation strategy. Pp. 85–106 in D. Sarewitz, R.A. Pielke Jr, and R. Byerly, editors. *Prediction: Decision Making and the Future of Nature.* Washington, DC: Island Press.

Evans, R. 1999. Economic models and policy advice: Theory choice or moral choice? *Science in Context* 12: 351–376.

Hooke, W.H., and R.A. Pielke Jr. 2000. Short-term weather prediction: an orchestra in search of a conductor. Pp. 61–84 in D. Sarewitz, R.A. Pielke Jr, and R. Byerly, editors. *Prediction: Decision Making and the Future of Nature.* Island Press: Washington, DC.

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

———. 1992. Discussion of 'The modeling process and model validation' by Chin-Fu Tsang. *Ground Water* 30: 622–623.

Konikow, L., and J.D. Bredehoeft. 1992. Groundwater models cannot be validated. *Advances in Water Resources* 15: 75–83.

Konikow, L., and E.P. Patten Jr. 1985. Groundwater forecasting. Pp. 221–270 in M.G. Anderson and T.P. Burt, editors. *Hydrological Forecasting.* Chichester: John Wiley and Sons.

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

Konikow, L. and L.A. Swain. 1990. Assessment of predictive accuracy of a model of artificial recharge effects in the upper Cochella Valley, California. Pp. 433–439 in E.S. Simpson and J.M. Sharp, editors. *Selected Papers on*

Hydrology from the 28th International Geological Congress (1989). Vol. 1. Hannover: Heinz Heise.

Kuhn, T.S. 1957. *The Copernican Revolution*. Cambridge, MA: Harvard University Press.

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

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

Moore, S. 1995. The coming age of abundance. Pp. 110–139 in R. Bailey, editor. *The True State of the Planet*. New York: Free Press.

Morgan, M.S. 1999. Learning from models. Pp. 347–388 in M.S. Morgan and M. Morrison, editors. *Models as Mediators*. Cambridge: Cambridge University Press.

Morgan, M.S., and M. Morrison, editors. 1999. *Models as Mediators*. Cambridge: Cambridge University Press.

Murphy, A.H. 1978. Hedging and the mode of expression of weather forecasts. *Bulletin of the American Meteorological Society* 78: 371–373.

Narisimhan, T.N. 1998. Quantification and groundwater hydrology. *Ground Water* 36: 1.

Nigg, J. 2000. The issuance of earthquake 'predictions': Scientific approaches and strategies. Pp. 135–156 in D. Sarewitz, R.A. Pielke Jr, and R. Byerly, editors. *Prediction: Decision Making and the Future of Nature*. Washington, DC: Island Press.

O'Neill, R.V. 1973. Error analysis of ecological models. Pp. 898–908 in D.J. Nelson, editor, *Proceedings of the Third National Symposium on Radioecology*, No. 547. Environmental Sciences Division, Oak Ridge, TN: Oak Ridge National Laboratory.

Oreskes, N. 1998. Evaluation (not validation) of quantitative models. *Environmental Health Perspectives* 106: 1453–1460.

———. 2000a. Why believe a computer? Models, measures and meaning in the natural world. Pp. 70–82 in J.S. Schneiderman, editor. *The Earth around Us: Maintaining a Livable Planet*. New York: W.H. Freeman.

———. 2000b. Why predict? Historical perspectives on prediction in earth science. Pp. 23-40 in D. Sarewitz, R.A. Pielke Jr, and R. Byerly, editors. *Prediction: Decision Making and the Future of Nature*. Washington, DC: Island Press.

Oreskes, N., and K. Belitz. 2001. Philosophical issues in model assessment. Pp. 23–41 in M.G. Anderson and P.D. Bates, editors. *Model Validation: Perspectives in Hydrological Science*. Chichester: John Wiley and Sons.

Oreskes, N., K. Shrader-Frechette, and K. Belitz. 1994. Verification, validation, and confirmation of numerical models in the earth sciences. *Science* 263: 641–646.

Petterssen, S. 2001. *Weathering the Storm: Sverre Petterssen, the D-Day Forecast, and the Rise of Modern Meteorology*, J.R. Fleming, editor. Boston: American Meteorological Society.

Pielke, R.A. Jr. 1999. Who decides? Forecasts and responsibilities in the 1997 Red River Flood. *Applied Behavioral Science Review* 7: 83–101.

Pilkey, O.H. Jr. 1990. A time to look back at beach nourishment (editorial). *Journal of Coastal Research* 6: iii–vii.

———. 1994. Mathematical modeling of beach behaviour doesn't work. *Journal of Geological Education* 42: 358–361.

———. 2000. Predicting the behavior of nourished beaches. Pp. 159–184 in D. Sarewitz, R.A. Pielke Jr, and R. Byerly, editors. *Prediction: Decision Making and the Future of Nature*. Washington, DC: Island Press.

Rastetter, E.B. 1996. Validating models of ecosystem response to global change. *Bioscience* 46: 190–198.

Sarewitz, D., and R.A. Pielke Jr. 2000. *Breaking the Global-Warming Gridlock, Atlantic Monthly July 2000*. http://www/theatlantic.com/issues/2000/07/sarewitz.htm.

Sarewitz, D., R.A. Pielke Jr, and R. Byerly, editors. 2000. *Prediction: Decision Making and the Future of Nature*. Island Press: Washington, DC.

Shackley, S., P. Young, S. Parkinson, and B. Wynne. 1998. Uncertainty, complexity, and concepts of good science in climate change modeling: Are GCMs the best tools? *Climatic Change* 38: 159–205.

Tsang, Chin-Fu. 1991. The modeling process and model validation. *Ground Water* 29: 825–831.

———. 1992. Reply to the preceding discussion of "The modeling process and model validation." *Ground Water* 30: 622–624.