ELSEVIER

The Science of the Total Environment 183 (1996) 33–47

the Science of the Total Environment
An International Journal for Scientific Research into the Environment and its Relationship with Man

Testing and validating environmental models

James W. Kirchner*[a], Richard P. Hooper[b], Carol Kendall[c], Colin Neal[d], George Leavesley[e]

[a]*Department of Geology and Geophysics, University of California, Berkeley, California 94720-4767, USA*
[b]*U.S. Geological Survey, Atlanta, Georgia 30360-2824, USA*
[c]*U.S. Geological Survey, 345 Middlefield Rd., Menlo Park, California 94025, USA*
[d]*Institute of Hydrology, McLean Building, Crowmarsh Gifford, Wallingford, Oxon OX10 8BB, UK*
[e]*U.S. Geological Survey, Box 25046, MS 412, Denver Federal Center, Lakewood, Colorado 80225, USA*

Abstract

Generally accepted standards for testing and validating ecosystem models would benefit both modellers and model users. Universally applicable test procedures are difficult to prescribe, given the diversity of modelling approaches and the many uses for models. However, the generally accepted scientific principles of documentation and disclosure provide a useful framework for devising general standards for model evaluation. Adequately documenting model tests requires explicit performance criteria, and explicit benchmarks against which model performance is compared. A model's validity, reliability, and accuracy can be most meaningfully judged by explicit comparison against the available alternatives. In contrast, current practice is often characterized by vague, subjective claims that model predictions show 'acceptable' agreement with data; such claims provide little basis for choosing among alternative models. Strict model tests (those that invalid models are unlikely to pass) are the only ones capable of convincing rational skeptics that a model is probably valid. However, 'false positive' rates as low as 10% can substantially erode the power of validation tests, making them insufficiently strict to convince rational skeptics. Validation tests are often undermined by excessive parameter calibration and overuse of ad hoc model features. Tests are often also divorced from the conditions under which a model will be used, particularly when it is designed to forecast beyond the range of historical experience. In such situations, data from laboratory and field manipulation experiments can provide particularly effective tests, because one can create experimental conditions quite different from historical data, and because experimental data can provide a more precisely defined 'target' for the model to hit. We present a simple demonstration showing that the two most common methods for comparing model predictions to environmental time series (plotting model time series against data time series, and plotting predicted versus observed values) have little diagnostic power. We propose that it may be more useful to statistically extract the relationships of primary interest from the time series, and test the model directly against them.

Keywords: Ecosystem models; Model evaluation

1. The need for modelling standards

Mathematical models of environmental systems are widely used in important — and controversial — public policy decisions, such as: How will

* Corresponding author. Tel.: +1 510 6438559; fax: +1 510 6439980; e-mail: kirchner@moray.berkeley.edu.

Elsevier Science B.V.
SSDI 0048-9697(95)04971-3

changes in greenhouse gas emissions affect Earth's climate? How will changes in climate alter the structure and function of natural ecosystems (or managed ecosystems, such as those underlying agriculture and forestry)? What reduction in 'acid rain' precursors would be needed to preserve or restore water quality in acid-sensitive regions? Are salmon fisheries collapsing due to natural population fluctuations, overfishing at sea, damming and diversion of spawning rivers, or the effects of land use on headwater streams? Will a proposed mitigation plan successfully contain groundwater contaminant plumes at a Superfund site? What is the probabilility that stored radioactive wastes will escape, over the next 10 000 years, under various repository schemes? These hotly contested public policy questions share several important characteristics. None are amenable to direct experimentation, none can be decided by simple extrapolation from past experience, and none concern systems that are simple enough that human intuition — even the intuition of experts — can provide a reliable guide for action.

Computer simulation models may provide useful insight into such problems, particularly if they are at least as reliable as the next best alternative (such as expert opinion). Although model reliability is a major factor in determining how models should be used in decision-making, tests of model reliability rarely receive the attention or emphasis they deserve. One reason, no doubt, is the sheer effort required to get complex models up and running. Another reason may be the lack of data sets with adequate spatial and temporal resolution. However, we suspect that many models are not rigourously tested simply because the community does not demand it. Instead, many journals routinely publish papers in which the authors simply opine that the model 'provides acceptable agreement with the data', without specifying their criteria for deciding what's acceptable, without objectively measuring how good the agreement actually is, without considering whether better agreement could be obtained from other models, and without addressing whether good agreement could have been obtained even if the model were fundamentally flawed.

Generally accepted standards for model evaluation are needed to encourage the development of better models. There will be little incentive to improve a model, unless there is also systematic pressure to find out how good or bad it actually is. Rigourous model evaluation, rigourously applied, would have the salutary effect of encouraging and rewarding better modelling efforts.

Model evaluation standards are also needed to reassure those who are skeptical of models and modelling. Some skeptics regard models as reckless exercises in 'garbage in, garbage out' that confer an air of scientific sophistication on unsubstantiated conjectures. That such models have sometimes been built is clear; whether they are the exception or the rule remains an open question. This question can only be resolved if there are generally accepted standards for distinguishing good models from bad, and if the community demands that these standards be applied.

Model evaluation standards are clearly desirable, but are they possible? The term 'mathematical modelling' embraces many diverse approaches, developed within many different disciplines, serving many different objectives, and using many different kinds of data. Any particular testing procedure might be ideal for one modelling approach, but inappropriate for others.

However, we believe there is a middle path between the current laissez-faire attitude that says, 'anything goes', and a methodological straightjacket that ignores critical differences among models. We argue that the longstanding scientific traditions of disclosure and documentation provide a useful guide for what we should expect of modellers, without constraining them to specific procedures, criteria, and protocols. The generally accepted standards for reporting research results do not require that all experiments be performed in a particular way. Instead, they simply require that all relevant factors should be disclosed and the basis for any conclusions should be documented. Analogously, we hold that modellers cannot be required to build their models according to a single method, or test them against a single criterion, but modellers can and should be required to disclose the tests that they have conducted (or disclose the fact that the model has not been tested at all). For example, if a model is

deemed to give 'acceptable agreement with the data', complete disclosure would include an assessment of how many adjustable parameters there are, how much flexibility they introduce into the model, and how probable it is that the model could fit the data even if it had critical flaws. Likewise, modellers cannot be required to apply a single universal protocol to reaching their conclusions, but they can and should be required to document the basis for whatever conclusion was reached. Documentation for a conclusion would include, for example, an assessment of how the conclusion depends on the premises and data underlying the model (sensitivity analysis), and an evaluation of how uncertainties in those premises and data affect the conclusion (uncertainty analysis). By 'documentation' we expressly do not mean simply documentation of the computer code; we hardly discharge our responsibility to our colleagues and the public if we say, 'Here, it's all in the code, you figure it out'.

The principles of disclosure and documentation demand that we move beyond the arbitrary model acceptance criteria commonly used today. A criterion like 'acceptable agreement with the data' is too subjective to provide effective documentation of how good or bad the agreement really is, particularly for those who have different standards of acceptability. Yet model evaluations are usually phrased in vague terms such as 'surprisingly good agreement with the data', or 'acceptable for the purposes of this study'. We should be able to do better than this.

Assessments of value only have meaning with reference to some benchmark for comparison. Models are only 'realistic', 'reasonable', 'valid', 'accurate' and so forth, compared to some alternative; such statements have little meaning on an absolute basis. Possible benchmarks for comparison include null hypotheses, alternative mathematical models, or expert judgment. Are the model's predictions markedly more reliable than those made by flipping coins (a null hypothesis)? Are they more accurate than predictions made by other models, or by experts? If so, then at what cost? If not, then why model?

At a minimum, documentation of model evaluation requires three elements: a performance criterion, a benchmark, and an outcome. The performance criterion (e.g. ability to match short-term fluctuations in the data, or precise agreement with observed long-term averages) documents which aspects of the model were tested, and in what way model performance was deemed good or bad. The 'benchmark' is the alternative the model was compared to. Specifying the benchmark answers the question, 'good compared to what?' Finally, specifying the outcome documents how good model performance was (and particularly, how much better or worse than the alternative).

If a model is to be used for policy analysis, the most obvious and appropriate benchmarks are the decision tools that would otherwise be used (such as expert opinion). Models are widely used under the premise that many environmental systems are too complex for expert judgment to be reliable. On the other side are those who argue that models are primarily useful to educate experts' intuition, which should then be used for prediction and decision-making [1,2]. Either premise may well be true; we simply point out that they both can be tested. It may be that experts are actually less reliable than a model, but are less easily revealed as such, because erroneous predictions are revised in hindsight. Expert judgments are also usually less concrete and specific than model results; this may make them harder to falsify, but it also may more accurately reflect the uncertainties inherent in predicting the behavior of complex environmental systems.

2. The value of strict tests

It is almost axiomatic that models should be rigourously tested, but this belief (with which we agree) leaves important questions unanswered. What makes a particular test rigourous? How can we distinguish between rigourous tests and those that are not rigourous? Can we measure how rigourous a test is? And precisely why are rigourous tests desirable?

In this section, we argue that a test's rigour is most directly gauged by the probability that invalid models could nonetheless pass it. Tests for which this probability is small (which we term 'strict' tests) can be shown to convey significant

information about model validity. By contrast, passing non-strict tests conveys little information about model validity. Most importantly, only strict tests can forge consensus among individuals with widely differing preconceptions of a model's validity, because only strict tests can be shown to be capable of convincing an open-minded skeptic. In contrast, tests that are not strict cannot be expected to substantially alter individuals' preconceptions about model validity; skeptics can rationally remain skeptical, even if the model passes the test. We show semi-quantitatively that most model validation exercises are probably not strict enough to convey much information about model validity.

Consider a greatly simplified example of model validation. In this example, the model being tested is either 'valid' or 'invalid', with no shades of gray in between, and the test is assumed to have a decisive outcome: either the model is declared to have 'passed', or it is declared to have 'failed'. We concede that the real world is not so simple. In reality, of course, models are rarely either valid or invalid; instead, they are valid to varying and uncertain degrees, in various particular ways, for various particular purposes. (By 'validity' we mean adequacy for a specific purpose, rather than absolute truth in every respect. All models simplify reality and therefore are unrealistic to some degree.) Likewise, results of model tests are rarely an unambiguous 'pass' or 'fail'; usually the model's performance lies somewhere between complete success and complete failure. Nonetheless, these simplifications make the following discussion much more straightforward, and they could be relaxed if a more elaborate treatment were desired.

Two potential points of confusion need to be clarified before we proceed. First, in our example, we have specified that the model can only have two states ('valid' or 'invalid'), but we have not specified which state it is in (that is, the model's validity is uncertain). Second, although we assume that the test results are unambiguous, this does not imply that the validity of the model is similarly unambiguous. That is, passing the test does not automatically imply that the model is valid, nor does failing the test automatically imply that the model is invalid. This would only be true for a perfect test, one that the model would pass if and only if it were valid. No tests are perfect. Instead, in the real world an invalid model will sometimes pass a test (by simple random chance, for example, or if the model's invalid features are masked by parameter tuning), and likewise a valid model will sometimes fail a test (if, for example, the data are unrepresentative). In our example, as in the real world, even though it might be clear whether the model has passed or failed the test, it might still be unclear whether the model is valid or invalid.

There will always be uncertainty surrounding whether a model is valid or not; the next few paragraphs explore how model tests can reduce this uncertainty. The analysis outlined below uses a probabilistic (or Bayesian) framework [3,4]. If readers are unfamiliar with Bayesian inference, we ask them to bear with us for a moment, since they may ultimately find this approach more intuitively appealing than the classical models for scientific inference.

We use $P(valid)$ to represent our confidence that the model is valid, before the test results are known, and use the conditional probability $P(valid|pass)$ to express our confidence that the model is valid, given that the model has passed the test. Although these levels of confidence are expressed in the formalism of probabilities, and although they can be manipulated according to probability theory, they are not probabilities in the classical sense. In classical scientific inference, it makes no sense to talk about the 'probability' that a model is valid, since validity is a simple question of fact; either the model is valid or it isn't. Most practicing scientists, however, find it natural to speak of the likelihood that a theory (or model) is valid; in this way they express their confidence or certainty, rather than making an objective statement of probability in the classical sense. Mirroring this habit of thought, the Bayesian approach does not formally distinguish between uncertainty and improbability. Thus, the explicitly Bayesian approach used here does not propose a new and different logic for scientific inquiry. Instead, it only makes practicing scientists' intuitive logic explicit, in part to explain why this logic is reasonable.

So, what is the probability that the model both (a) is valid, and (b) passes the test? This can be calculated in two ways. One is the pre-test probability (or confidence) that the model is valid, times the probability that the model will pass the test if it is valid. The second is the probability that the model will pass the test, times the probability that it is valid if it passes the test. These two probabilities must be equal, so

$$P(valid|pass) \times P(pass) = P(pass|valid) \times P(valid) \quad (1)$$

Eq. (1) can be rearranged to yield the Bayesian updating rule,

$$P(valid|pass) = \frac{P(pass|valid)}{P(pass)} P(valid) \quad (2)$$

Eq. (2) shows that our confidence $P(valid|pass)$ that the model is valid (given that it has passed the test) depends on $P(valid)$, our confidence before the test, times a ratio that describes the characteristics of the test, namely, the ratio between $P(pass|valid)$ (the probability that a valid model would pass) and $P(pass)$ (the probability that the model would pass whether or not it is valid). We can expand $P(pass)$ into its two components,

$$P(pass) = P(pass|valid) \times P(valid) + P(pass|invalid) \times [1 - P(valid)] \text{ where}$$

$$1 - P(valid) = P(invalid) \quad (3)$$

Where $P(pass|valid)$ is the probability that a valid model would pass the test, and $P(pass|invalid)$ is the probability that an invalid model would pass. Although these probabilities may be difficult to estimate for any particular test, they are not subjective like our pre-test confidence, $P(valid)$. The Bayesian updating rule thus becomes:

$$P(valid|pass) = \frac{1}{1 + \frac{P(pass|invalid)}{P(pass|valid)} \left[\frac{1}{P(valid)} - 1\right]} \quad (4)$$

or, if the model fails the test,

$$P(valid|fail) = \frac{1}{1 + \frac{P(fail/invalid)}{P(fail/valid)} \left[\frac{1}{P(valid)} - 1\right]} \quad (5)$$

Eqs. (4) and (5) are rational rules for using model tests to revise our prior assessments of a model's validity. Both expressions unavoidably depend on *P(valid)*, the subjective pre-test confidence that the model is valid. Although *P(valid)* is subjective, the relationships linking *P(valid)*, *P(valid|pass)*, and *P(valid|fail)* are neither subjective nor arbitrary. Given a specified pre-test confidence, the post-test assessments of validity are not arbitrary, but instead are completely determined by the outcome of the test and the conditions under which that result was obtained, that is, the probabilities *P(pass|valid)* and *P(pass|invalid)*. In this sense, Eqs. (4) and (5) are rational, even though they have a subjective element. Furthermore, although *P(valid)* is subjective, properly designed tests can make post-test assessments of validity — *P(valid|pass)* and *P(valid|fail)* — considerably less subjective. As we will shortly show, for sufficiently strict tests, *P(valid|pass)* is largely independent of pre-test confidence in model validity.

The pre-test confidence in the model's validity, *P(valid)*, will vary among individuals. For a true believer, such as a model's author, *P(valid)* may be high (say, 0.9), while for skeptics, such as the author's competitors, *P(valid)* may be low (say, 0.1), and an impartial observer, with no basis for making a prior judgment, might assume that *P(valid)* is 0.5. Assume that the skeptic is not so dogmatic as to assume that *P(valid)* is exactly zero, and the true believer is not so dogmatic as to assume that *P(valid)* is exactly one. Now, what test could convince all three individuals that the model is probably valid? Inspection of Eq. (4) reveals that post-test confidence in the model can approach 1, independent of the subjective pre-test confidence *P(valid)*, if the test is sufficiently strict, that is, if it is sufficiently unlikely that an invalid model would pass.

As Fig. 1 shows, a test that is not strict will barely alter the preconceptions of the three individuals; the believer remains enthusiastic, and the skeptic remains skeptical. A test's capacity to alter prior assessments of validity is a measure of how much information it provides. From this perspective, tests that are not strict cannot convey much information about model validity. Only strict tests are able to create consensus between believers and skeptics.

The degree of rigour in a test is commonly viewed as simply 'how hard it is to pass'. However, one

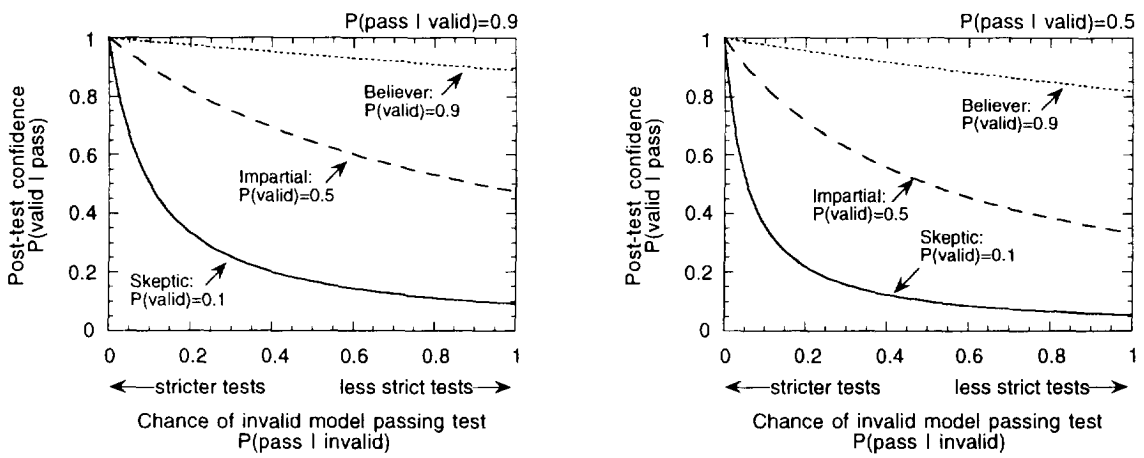

Fig. 1. Assessments of model validity after a test has been passed, *P(valid|pass)*, calculated from Eq. (4), as a function of the test's strictness, for individuals with three different prior assessments of model validity: a 'believer', a skeptic, and an impartial observer (who believe the chances that the model is valid are 0.9, 0.1 and 0.5, respectively). The test's strictness is expressed by the false positive rate, *P(pass|invalid)*, which is the likelihood that an invalid model would pass the test. Tests that are not strict (tests for which *P(pass|invalid)* is not small) do little to alter the three individuals' subjective assessments. However, sufficiently strict tests (tests that invalid models are unlikely to pass) force all three individuals to agree that the model is likely to be valid, despite their prior disagreements. False positive rates must be low to forge consensus between believers and skeptics, whether the tests are easy for models to pass (*P(pass|valid)* = 0.9, left panel), or more difficult for valid models to pass (*P(pass|valid)* = 0.5, right panel).

can see from Fig. 1 that increasing the difficulty of the test for valid models (changing *P(pass | valid)* from 0.9 to 0.5) does little to alter the inference that would be drawn if the model passes. If a model has passed the test, Fig. 1 and Eq. (4) show that rational post-test assessments of model validity have little to do with how difficult the test was for a valid model to pass, but they have a lot to do with how difficult it would be for an invalid model to pass. That is, the test's difficulty is less important than its selectivity (its 'false positive' rate). In other words, successful test results are meaningful only if an invalid model could probably not have passed. Put in such direct terms, this general principle should be intuitively obvious to most practicing scientists. Thus the preceding analysis has not discovered a new principle, but it has explained why strict tests are valuable, and has described precisely what makes strict tests strict.

But how strict is strict enough? Eq. (4) allows us to quantify how strict a test must be in order to convey significant information. The results are disquieting. As shown above, the information content of a test depends on the 'false positive' rate, the probability that an invalid model could pass. It is admittedly difficult to quantify this probability. However, our assessment (relying on our combined experience both in modelling, and in evaluating models built by others) is that typical model validation exercises are probably not more than 80 or 90% effective in detecting invalid models. In other words, we believe that false positive rates could easily be at least 0.1 or 0.2, and possibly higher in some cases. However, as Fig. 1 clearly shows, false positive rates must be much lower than 0.1 or 0.2 for positive results to convince rational skeptics, and thus to create consensus between skeptics and true believers. In other words, our analysis provides a quantitative basis for concluding that typical model validation exercises are probably not strict enough to convey much information about model validity.

Our analysis demonstrates that the significance of a model test depends crucially on the false positive rate, or, the probability that invalid models could pass the test. Because the false positive rate plays such a central role, methods for quantifying it are urgently needed. One possibility is to generate synthetic data sets from several fundamentally incompatible models, then ask how often each model would 'fail', when tested against data generated from the other models. Here, although we will not attempt to quantify false positive rates, we will briefly comment on several factors that affect the likelihood of false positives.

The false positive rate depends on both the testing procedures and the model's characteristics. Parameter calibration is one feature of modelling practice that clearly could contribute to false positives, by masking deficiencies in model structure. Successful parameter calibration implies either that the model structure and the parameter values are both realistic, or that they are both unrealistic but compensate for one another. Particularly when the validation data set is functionally equivalent to the calibration data (as is often the case with two time periods from the same data series), parameter tuning can dramatically increase the false positive rate.

Parameter tuning makes the model more flexible, and reduces the degree to which the underlying model structure constrains model behavior. If model structure has little effect on model behavior, flaws in model structure are unlikely to be revealed by testing model behavior against data. In extreme cases, where there are too many free parameters, model calibration can become a computationally sophisticated — but scientifically empty — exercise in multidimensional nonlinear curve fitting. How many free parameters are too many? The answer remains unclear, but the available evidence indicates that many models have far more free parameters than can be reliably estimated from typical environmental data sets. For example, Jakeman and Hornberger [5] indicated that typical rainfall-runoff data contain only enough information to constrain simple hydrologic models having up to four free parameters. Hooper et al. [6] showed that even detailed hydrologic and geochemical time series data were insufficient to constrain a simple model with only six free parameters. In short, when it comes to parameter calibration, very little is still too much.

A related practice that also undermines the power of model tests is the use of ad hoc model features to make the model fit the data better. To take just one example, the Birkenes model [7] assumes that weathering reactions (which consume

H^+) will cease when $[H^+]$ decreases to 5 μM, but this assumption was not justified on the basis of weathering studies; instead, it was motivated simply by the observation that in the field data, $[H^+]$ never fell below 5 μM. Like parameter tuning, ad hoc empiricisms undermine the significance of a model test (by making spurious agreement between model and data more likely), but because they are put in during model development rather than as part of explicit calibration procedures, their implications for model testing can easily be overlooked.

Whether a model test is strict depends not only on the characteristics of the model being tested, but also on the nature of the data used to test it. Models are often tested against environmental time series, in which the relevant signals may be hidden by relatively large noise. In such situations, many different models may appear to fit the data equally well, within the scatter of the data. Thus the power of the test is low, because the data do not permit discrimination among alternative models.

One approach in such cases is to statistically extract the relationship of interest from the data set, and test the model against the statistically clarified relationship rather than the noisy raw data (see below). Another approach is to use chemical or isotopic tracers to specifically target the processes of interest. A third approach is to amplify the behavior of interest, through experimental manipulation of field plots. Finally, one can use controlled laboratory experiments to elucidate the behavior of interest, without the confounding factors that may dominate the data in nature. For example, laboratory-scale experiments are much less complex than real-world watersheds; their material properties can be much better characterized, and their behavior can be much more precisely and comprehensively monitored. Controlled experiments provide a more sharply defined empirical 'target', making it easier to tell if a model has missed the mark.

Laboratory experiments are rarely faithful models of the real world, but it would be hard to claim that although a particular model could not predict the outcome of a simple laboratory experiment, it nonetheless could reliably predict the behavior of real ecosystems, with their vastly greater complexity. Thus, while laboratory experiments may not provide strong confirmation for models, they have considerable diagnostic power. If controlled experiments focus on mechanisms that are relevant to the real-world problems of interest, they can provide strong disconfirmation, under the presumption that models that do not work under simplified laboratory conditions probably will not work in the real world either.

3. The value of relevant tests

For a test to be useful, it must be more than merely strict; it also must be relevant to the conditions under which the model will need to function. Models are usually intended to extrapolate beyond the range of historical experience. It does little good to test such a model against time series that exhibit the same range of behaviors as the calibration data. Instead, it makes more sense to test the model against data that diverge substantially from the calibration. Experimental ecosystem manipulations can make such data available, and can be designed to focus on processes or forcing factors of particular interest. For example, Wright et al. [8,9] experimentally altered acid loading to several small catchments, producing marked changes in runoff chemistry; they then tested an acidification model against the resulting runoff data. Long-term monitoring data can also reveal fortuitous changes in natural or anthropogenic forcing. For example, Kirchner et al. [10] used climatically triggered acid episodes to test a theory of catchment acid buffering. The acid episodes produced dramatic departures from pre-episode stream chemistry; in addition, because the theory did not permit calibration in the normal sense, the test was unusually strict.

Models are often needed to predict ecosystem response to particular kinds of forcing, over particular time scales. If so, tests emphasizing other types of forcing, or other time scales, may not be relevant. Even if a model passes a strict test, it does not follow that all aspects of that model have been tested. At best, test results can assess only those parts of the model that respond to the imposed forcing, and affect the observed output variables, under the test conditions.

Time series data from natural ecosystems (e.g.

chemical concentration and water flux time series from catchment studies, measures of species abundance in ecological studies, or weather observations in climate modelling studies) are frequently used in model testing. However, environmental data reflect many different types of forcing, some of which are more relevant than others for the policy decisions that must be made. Again taking catchment acidification models as an example, we note that catchment runoff chemistry may be affected by at least three factors: hydrologic fluctuations, changes in atmospheric deposition, and long-term depletion of base cations from catchment soils. The processes of greatest policy interest are the direct effects of atmospheric deposition on runoff chemistry, and the indirect effects of acid loading on base cation depletion from soils. Some hydrochemical models [7,11] explicitly model the short-term effects of catchment hydrology on runoff chemistry. If these models are tested against data that are dominated by hydrologic fluctuations rather than chemical changes, they might correctly predict chemical response to the storm hydrograph, and thus explain most of the observed variance, even if they did not correctly predict catchment response to chemical forcing (see below).

Most acidification models are also designed to predict how runoff chemistry will respond to long-term depletion of base cations from catchment soils, under accelerated leaching by acid deposition. However, available catchment data do not clearly reveal the effects of changes in base saturation, because this process may take decades to produce measurable effects, and because these trends may be obscured by short-term changes in catchment hydrology, atmospheric forcing, and other factors. Therefore, model predictions of chronic acidification from base cation depletion are still largely untested. Because these long-term predictions have far-reaching public policy implications, there is an urgent need for tests that focus on the effects of base cation leaching. Because the relevant processes take place too slowly in nature, these tests may only be possible with data from artificially accelerated laboratory experiments [12].

4. Testing models against environmental time series

Models are often tested against time series data from natural ecosystems. Time series data can be compared to models by several methods. Below, we show that two common methods may fail to reveal significant discrepancies between model and data. We argue that it is often useful to isolate and clarify the relationships among variables of interest in the data, before comparing these relationships to model predictions.

These points are illustrated below, using synthetic time-series data from a hypothetical watershed (Fig. 2). Using synthetic data permits us to know what the underlying 'true' model is, because we have specified it. In other respects, however, the data examined here are similar to real-world catchment data recently analyzed by Kirchner et al. [13]; the distributions of each variable are similar to those found in real-world catchment data, and each variable exhibits realistic levels of correlation with the others, as well as realistic levels of serial correlation with itself.

The problem at hand is to predict how changes in sulfate concentrations will affect acid neutralizing capacity (ANC), in order to evaluate the expected benefits from reductions in sulfate loading. ANC is also expected to be a function of catchment discharge, but this relationship is less important for policy analysis purposes. Changes in discharge may also affect sulfate concentrations; thus, discharge can affect ANC both directly, and indirectly via sulfate.

Assume that we want to test two models that predict ANC. The first, 'model A', predicts that ANC should be a linear function of the log of discharge, and a linear function of sulfate concentration:

$$\text{Model A: } ANC = 213 - 80 \times \log(\text{flow}) - 0.4 \times SO_4 \quad (6)$$

The second model, 'model B', predicts that ANC will be independent of sulfate concentration, and will be affected slightly differently by discharge:

$$\text{Model B: } ANC = 91 - 67 \times \log(\text{flow}) \quad (7)$$

In Eqs. (6) and (7), the coefficients could be physically-based, or they could be determined by calibration to independent data; it makes no dif-

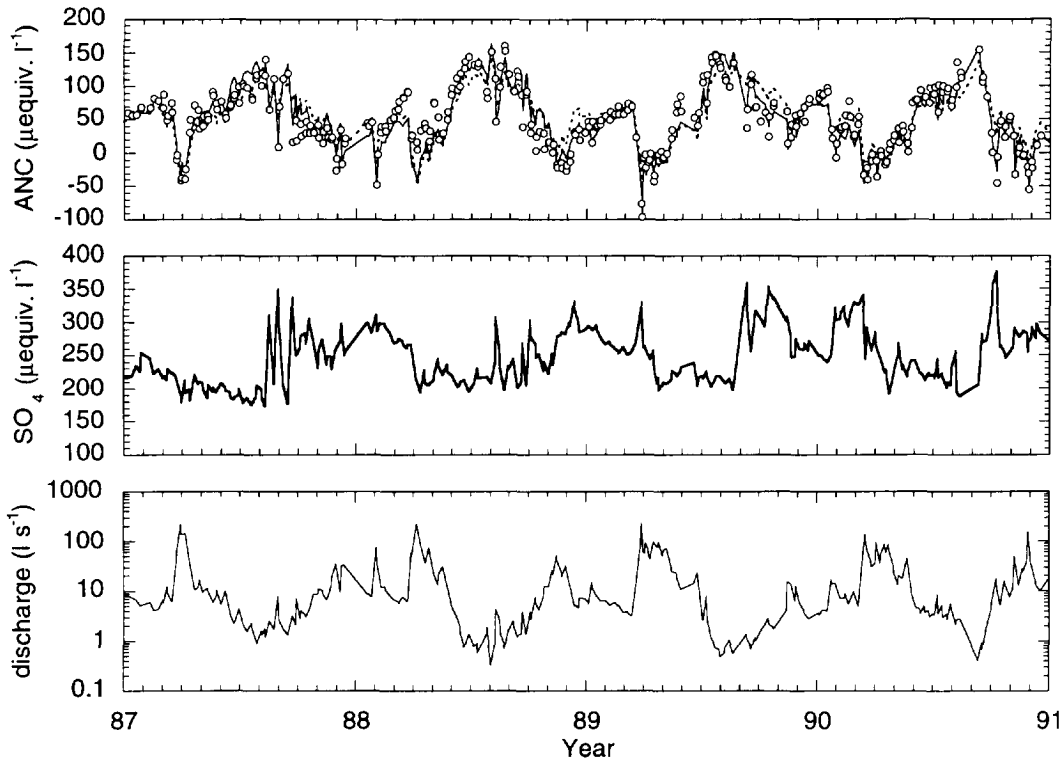

Fig. 2. Time series data for discharge, sulfate concentration, and acid neutralizing capacity (ANC) in a hypothetical catchment. Synthetic ANC data shown by open circles. ANC predicted by model A (see text) shown by solid line; ANC predicted by model B (see text) shown by dotted line.

ference for the purposes of the model tests shown below. We emphasize, however, that we are not concerned with fitting either model to the data. Instead, we take models A and B to be completely specified. Although they are linear equations, this is just to make the demonstration simple; they should not be mistaken for regression models. In actual practice, the models being tested might be complex sets of coupled equations, rather than the simple closed-form expressions shown here; this complicates the situation but does not fundamentally alter the principles involved.

Now we turn to the problem of testing model A and model B against the data. The most common

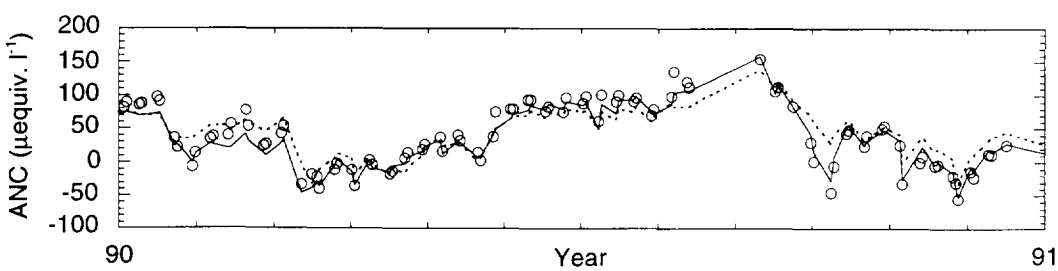

Fig. 3. Expanded view of 1 year from hypothetical ANC time series from Fig. 2, more clearly showing predictions from model A (——) and model B (– – –).

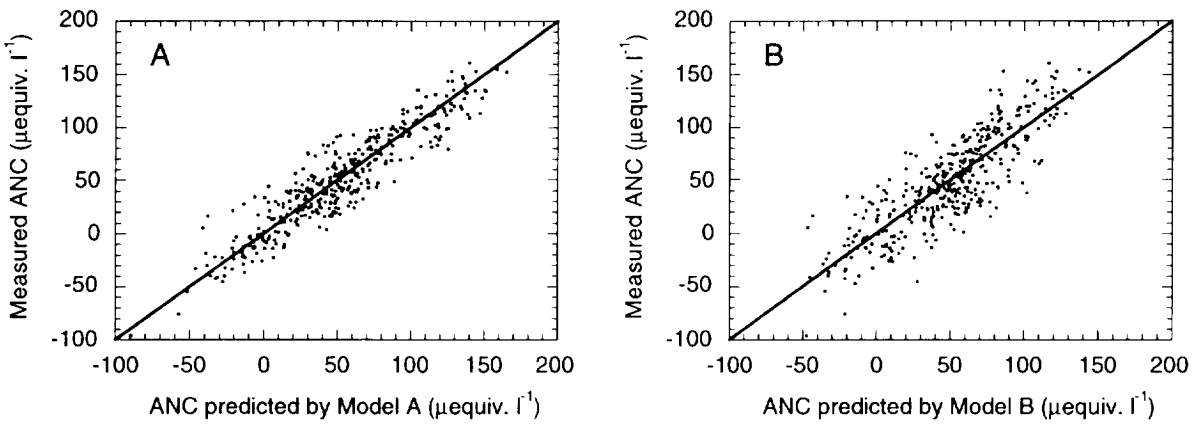

Fig. 4. ANC values predicted by model A (left panel) and model B (right panel), compared to ANC values in hypothetical time series. Diagonal line indicates perfect agreement between model and observations.

way that environmental models are compared with data is by visually comparing the time-trace of the model with the time series of the data. Examining Figs. 2 and 3, it is hard to find anything wrong with either model. Both models appear to do a good job of following the ups and downs of the ANC time series, within the apparent scatter of the data.

The second common method for comparing models to data is by plotting the observed values as a function of model predictions, and visually checking whether the data conform to the line of perfect agreement. Testing models A and B in this way (Fig. 4) reveals no systematic discrepancies between either model and the data. Although model B appears to predict ANC slightly less accurately than model A, neither model shows visible bias. If the two models were explicitly compared, model A might appear preferable, but if only one model were tested (as in conventional practice), either would appear to predict ANC accurately.

So, using the most common methods for comparing model predictions against data, we find no reason to reject either model. But models A and B cannot both be accurate, because they are inconsistent with one another; model A assumes that sulfate depresses ANC by a specified amount, while model B assumes that there is no such effect. One (or possibly both) must be wrong, even though both fit the ANC data well. Despite the difference between the two models — a difference that is crucial for policy analysis — either model looks 'acceptable'. Is there any way to discriminate between the models?

To better visualize how SO_4 affects ANC, and thus better evaluate the two models, one could plot ANC directly as a function of SO_4, as in Fig. 5 (left panel). The advantage of viewing the data in this coordinate space is that the difference between the predictions of the two models can be clearly seen. Unfortunately, the data show no clear pattern, and either model A or model B appears to describe the data equally well (or perhaps equally poorly). The regression line through the data (shown as a dashed line in Fig. 5) lies between model A and model B, and arguably represents the best estimate of whatever linear trend is present. The regression slope is −0.2, with a standard error of 0.05, and thus it differs from both model A (slope of −0.4) and model B (slope of zero) by four times its standard error. In a narrow statistical sense, then, Fig. 5 arguably invalidates both models. Model A apparently overstates the effect of SO_4 on ANC, while model B apparently underestimates it. In Figs. 2, 3 and 4, both models seemed to be valid, but now neither appears adequate. What has happened?

Because ANC depends on both SO_4 and discharge, either factor can obscure the effect of the other. In our example, the influence of discharge on ANC (Fig. 5, right panel) is two or three times

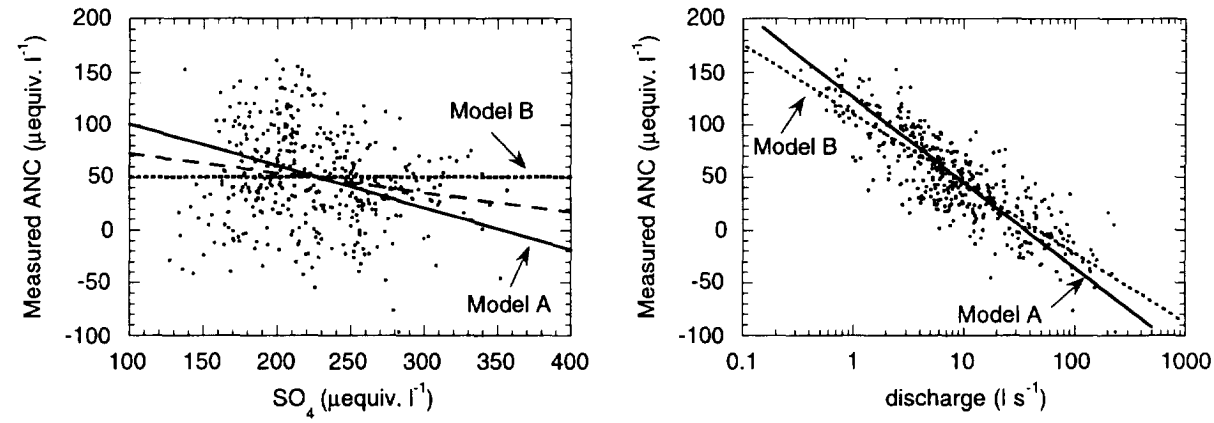

Fig. 5. ANC for hypothetical data set as a function of the two causal factors, sulfate concentration and discharge. Plotting ANC as a function of sulfate without correcting for flow variations (left panel) reveals no clear trend. Regression line is shown by broad dashes for comparison. A clearer relationship is evident between ANC and flow (right panel).

stronger than the effect of SO_4. This has two important consequences. First, any model that captures ANC's discharge-dependence will predict ANC relatively accurately, whether or not it correctly models the effect of SO_4, because that effect is a relatively small part of the ANC signal. That is why, even though the two models disagree over how to treat SO_4 effects, both models appear to fit the ANC data well. Second, because the large effect of discharge obscures the much smaller effect of SO_4, there is little apparent pattern in the left panel of Fig. 5. In these circumstances, simply plotting ANC against SO_4 may not reveal the functional dependence of ANC on SO_4. Although the effect of SO_4 is relatively small, we cannot simply dismiss it as unimportant; indeed, in our example, SO_4 is the only variable of practical importance for policy purposes.

It is tempting — but incorrect — to conclude that since there is no clear pattern between SO_4 and ANC, SO_4 must have no clear effect on ANC. To reveal the effect of SO_4 on ANC, and thus finally decide whether either model is likely to be adequate, we need to statistically separate the effects of SO_4 and discharge. In this case, because these effects are roughly linear, this separation can be done by multiple regression. (In more realistically complex cases, other techniques may

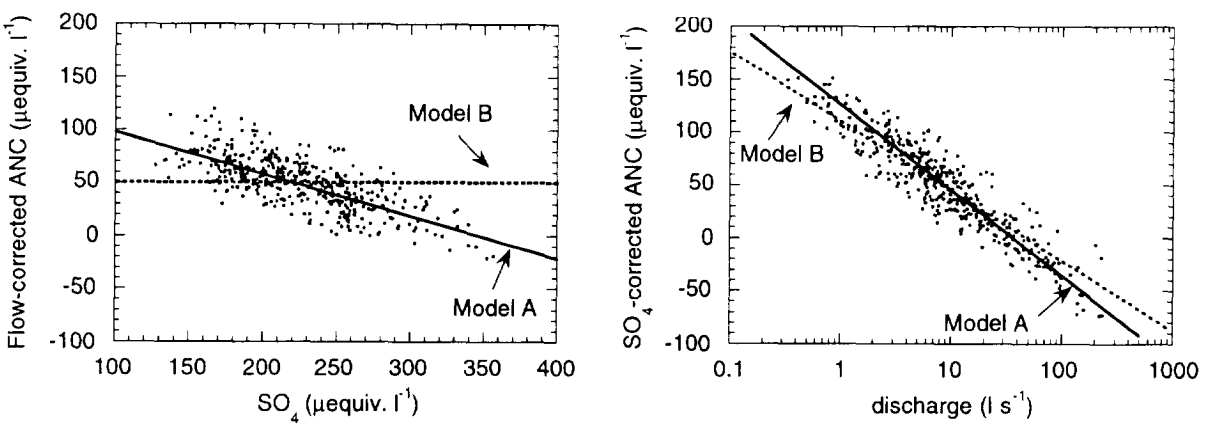

Fig. 6. Dependence of ANC on sulfate concentration and discharge, revealed by multiple regression. Removing flow effects clarifies the relationship between ANC and sulfate (left panel), revealing that model A is consistent with the data and model B is not. Correcting for changes in sulfate also clarifies ANC's dependence on flow (right panel).

be needed; see Kirchner et al. [13] for a slightly more complicated real-world example.) What is important is not the particular technique but the general principle of filtering out confounding factors to better reveal the interactions of greatest interest.

Applying multiple regression to our synthetic data, and correcting for the confounding effect of discharge variations, reveals the relationships shown in Fig. 6. Removing the scatter caused by discharge fluctuations reveals that the dependence of ANC on SO_4 is clear, in marked contrast to Fig. 5. The multiple regression slope of ANC on SO_4 is -0.4 ± 0.02, which is consistent with model A but not with model B. Model B does not accurately reflect the effect of SO_4 on ANC (the effect most relevant for policy analysis), but this was only revealed by statistically isolating that effect and correcting for the confounding effect of discharge. Careful readers will note that the multiple regression slope of the SO_4-ANC relationship here is twice as steep as the simple regression slope obtained in Fig. 5. The difference arises from the fact that SO_4 is negatively correlated with discharge in the data. Thus high flows, all else equal, tend to produce both low ANC and low SO_4; this partially masks the direct effect of low SO_4, which is to raise ANC. In other words, because SO_4 and discharge are correlated, Fig. 5 is a distorted — not merely blurred — representation of SO_4's causal effect on ANC.

The multiple regression slope of ANC on the log of discharge is -74 ± 2 μequiv. l^{-1} log $unit^{-1}$, which is significantly different from both models (Fig. 6). We could conclude, therefore, that both models have been shown to be invalid. However, while the difference between this regression slope and the two models is statistically significant, it is not practically significant, particularly since discharge is not a variable of policy interest. Thus model A appears to be adequate (although strictly incorrect), while model B appears to be inappropriate for our purposes.

Although model B incorrectly specified the mechanism of interest, only our last method of analysis revealed this to be the case. Evaluated by the other methods, model B appeared adequate. In particular, visual inspection of the time series (Figs. 2,3) and the plots of predicted versus measured ANC (Fig. 4) failed to detect any grounds for concern, yet these are precisely the methods most often used to evaluate model predictions. Only when the relationship of interest (ANC's dependence on SO_4) was extracted from the other confounding factors in the data, did it become clear that model B was fundamentally incorrect.

Another way to expose the difficulty with model B is to plot the difference between predicted and observed ANC as a function of sulfate (Fig. 7). These residuals of model B show a systematic pattern, indicating that the SO_4-dependence of ANC

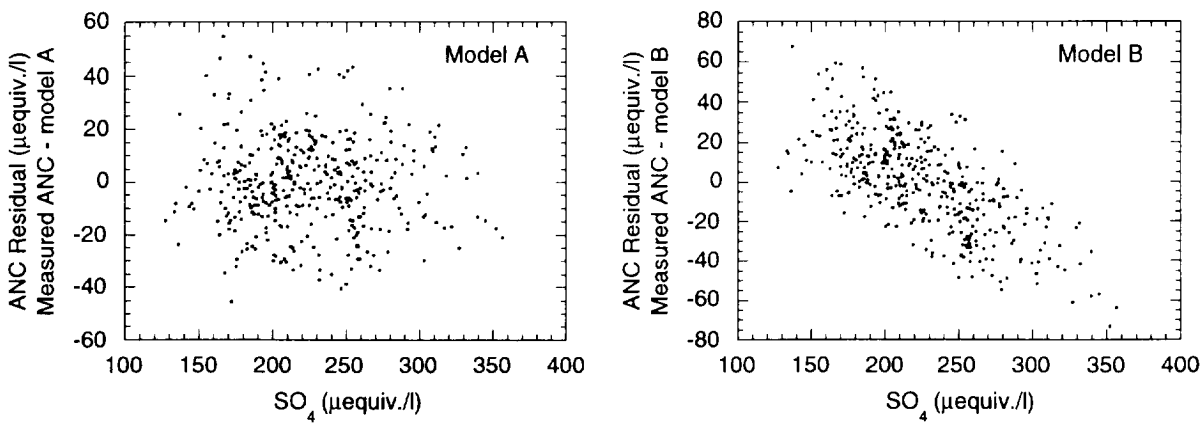

Fig. 7. Model residuals (observed ANC minus predicted ANC) plotted as a function of sulfate concentration. Residuals for model B (right panel) show clear dependence on sulfate, indicating a systematic difference between the responses of the model and the data to changes in sulfate. Residuals for model A (left panel) show no systematic deviation of model from data.

is different in the model and the data. In contrast, the residuals of model A show no systematic relationship with SO_4. Inspection of residuals is a standard step in evaluating statistical models, but is applicable to simulation modelling as well. Residuals reveal useful information about simulation models, just as they do in purely statistical analyses, by highlighting discrepancies between model and data.

This example was, we concede, contrived to demonstrate our point. For example, the effects of sulfate on ANC would be clearer in the data if sulfate varied more, or if flow varied less, or if the dependence of ANC on sulfate were not distorted by the correlation between sulfate and flow. However, the real world is rarely so convenient. Our data set is not a rare pathological case. On the contrary, it simply exhibits several common features (such as correlation between nominally independent variables) that often complicate environmental analyses. If anything, our data set is unrealistically straightforward, since the level of spurious noise is relatively low, the distributions are not highly skewed, and there are no outliers.

The reader might object that any puzzle is easy once one knows the answer, and might question whether we could have reached the correct answer if we had not known it beforehand. We share this concern, and it proves our point. Even in such a simple case as this — where there are only three variables, the underlying 'real world' relationships are linear, and lots of data are available — it can be exceedingly difficult to determine whether a given model accurately portrays the mechanisms behind the data. The task becomes immeasurably harder if there are dozens of potentially relevant variables, if these variables are interrelated in complex nonlinear ways, and if the data are sparse or unreliable. Because the most common model evaluation techniques utterly failed in our simple test case, we think they are unlikely to be reliable when applied to the more difficult task of real-world model testing.

Based on our combined experience, both with simplified test cases and with real-world modelling exercises, we believe that there are generalizable lessons to be learned from this brief demonstration. First, even in simple cases it can be very hard to tell whether the mechanisms in a model are realistic. Serious model flaws can be fiendishly difficult to uncover. As a result, model tests are rarely strict; the probability that flawed models will pass is rarely low.

Second, the conventional visual comparisons of time series (Figs. 2,3), and plots of predicted versus observed values (Fig. 4), can be singularly ineffective at revealing problems with models, even models as simple as the two tested here. Simply testing whether a model makes accurate predictions may not be very informative, if you need to know not only whether it gives the right answers, but also whether it does so for the right reasons.

Third, testing models against time series is difficult because the forcing factor of interest is often confounded with other factors, including irrelevant ones. It may often be more productive to extract the relationship of interest from the time series; that is, to isolate the relationships between the forcing factors and outcome variables of interest, correcting (insofar as possible) for the influence of other confounding factors (Fig. 6). This procedure should provide a stricter test of model structure, and one that is more relevant to the practical concerns behind many ecosystem models.

5. Conclusions

Improving ecosystem models will require setting higher standards for model testing and evaluation. An important first step, in our view, is to ask modellers to use explicit performance criteria in evaluating their models, and to compare them against explicitly stated benchmarks. This would be a significant improvement over the subjective model evaluations that are common today. Explicitly testing models against other decision-making methods (such as expert opinion) would provide a particularly illuminating measure of the accuracy and reliability of model predictions.

There is an urgent need for model tests that are sufficiently strict, tests that invalid models are unlikely to pass. To convince rational skeptics, validation tests must have very low false positive rates (Fig. 1). Our simple analysis shows that typical model tests are probably not strict enough to convince rational skeptics, and therefore are not strict enough to forge consensus on public policy issues. The power of model tests to detect flawed

models can be substantially eroded by parameter calibration and ad hoc model features. The power of model tests is also often limited by the available data, because in typical environmental time series, the signals of interest are often weak compared to the background noise. In such cases, tracer studies, plot-scale field manipulations, and laboratory experiments may provide more exacting tests. Cleverly designed experiments and tracer studies can emphasize the particular forcing factors of greatest interest, while minimizing the effect of potentially confounding factors. When environmental monitoring data are used to test models, clever applications of statistical methods can correct for confounding factors and, as much as possible, highlight the relationships between the relevant forcing factors and outcome variables.

Many of the conventional methods for testing models have relatively low power to detect serious flaws. A wide range of more powerful techniques is available, and they have recently come into limited use. These procedures will become more widely used if the community demands them. A clear consensus among modellers and model users demanding more exacting model tests could spur the development of better models and thus advance ecosystem science.

Acknowledgements

This work was supported, in part, by NSF grant EAR-9357931 to J.W.K. We thank M.E. Power and B.A. Roy for their comments on earlier versions of the manuscript.

References

[1] L.F. Konikow and J.D. Bredehoeft, Ground-water models cannot be validated. Adv. Water Resour., 15 (1992) 75–83.

[2] N. Oreskes, K. Shrader-Frechette and K. Belitz, Verification, validation, and confirmation of numerical models in the Earth sciences. Science, 263 (1994) 641–646.

[3] R.D. Rosenkrantz, Inference, Method, and Decision: Towards a Bayesian Philosophy of Science, Reidel, Boston, 1977.

[4] C. Howson and P. Urbach, Scientific Reasoning: The Bayesian Approach, Open Court, La Salle, IL, 1993.

[5] A.J. Jakeman and G.M. Hornberger, How much complexity is warranted in a rainfall-runoff model? Water Resour. Res., 29 (1993) 2637–2649.

[6] R.P. Hooper, A. Stone, N. Christophersen, E. de Grosbois and H.M. Seip, Assessing the Birkenes model of stream acidification using a multisignal calibration methodology. Water Resour. Res., 24 (1988) 1308–1316.

[7] N. Christophersen, H.M. Seip and R.F. Wright, A model for streamwater chemistry at Birkenes, Norway. Water Resour. Res., 18 (1982) 977–996.

[8] R.F. Wright, E. Lotse and A. Semb, Reversibility of acidification shown by whole-catchment experiments. Nature, 334 (1988) 670–675.

[9] R.F. Wright, B.J. Cosby, M.B. Flaten and J.O. Reuss, Evaluation of an acidification model with data from manipulated catchments in Norway. Nature, 343 (1990) 53–55.

[10] J.W. Kirchner, P.J. Dillon and B.D. LaZerte, Predicted response of stream chemistry to acid loading tested in Canadian catchments. Nature, 358 (1992) 478–482.

[11] S. Gherini, L. Mok, R.J.M. Hudson, G.F. Davis, C. Chen and R. Goldstein, The ILWAS model: formulation and application. Water Air Soil Pollut., 26 (1985) 95–113.

[12] R.A. Skeffington and D.J. Roberts, Testing a catchment acidification model: 'MAGIC' applied to a 5-year lysimeter experiment. J. Hydrol., 144 (1993) 247–272.

[13] J.W. Kirchner, P.J. Dillon and B.D. LaZerte, Separating hydrological and geochemical influences on runoff chemistry in spatially heterogeneous catchments. Water Resour. Res., 29 (1993) 3903–3916.