Available online at www.sciencedirect.com
SCIENCE DIRECT
ELSEVIER
Forest Ecology and Management 198 (2004) 89–103
Forest Ecology and Management
www.elsevier.com/locate/foreco

A synthesis of literature on evaluation of models for policy applications, with implications for forest carbon accounting

Stephen P. Prisley*, Michael J. Mortimer

Department of Forestry (0324), Virginia Tech, Blacksburg, VA 24061, USA

Received 29 December 2003; received in revised form 16 March 2004; accepted 18 March 2004

Abstract

Forest modeling has moved beyond the realm of scientific discovery into the policy arena. The example that motivates this review is the application of models for forest carbon accounting. As negotiations determine the terms under which forest carbon will be accounted, reported, and potentially traded, guidelines and standards are being developed to ensure consistency, accuracy, transparency and verifiability. To date, these guidelines have focused on definitions, data, and reporting, not models. The goal of this paper is to synthesize literature that may inform the development of guidelines for the application of models in areas with policy implications, such as forest carbon accounting. We discuss validation, verification, and evaluation as applied to modeling, and review common components of model evaluation. Peer review, quantitative analysis of model results, and sensitivity analysis are the most widely used approaches to model evaluation. US judicial and legislative perspectives on criteria for model acceptability are summarized.

Keywords: Forest models; Validation; Verification; Sensitivity analysis

1. Introduction

Models originating in scientific disciplines from atmospheric deposition to groundwater hydrology have been appropriated for use in developing policies and implementing regulations. When models are so used, there is an increased need to assess and communicate the suitability of the models for the specified tasks. The application that has motivated this paper is the use of forest carbon models to estimate and report carbon stock changes pursuant to provisions of international climate change agreements.

The Kyoto Protocol to the Framework Convention on Climate Change (UNFCCC, 1997) has elevated the role of forests in the global carbon cycle to a key component in international negotiations (Murray et al., 2000). Because of provisions in the Kyoto Protocol, forest carbon has the potential to become an economic commodity that may be exchanged in markets such as the Chicago Climate Exchange (Economist, 2002). There is now a compelling worldwide interest in being able to quantify and report forest carbon storage and flux in a transparent, verifiable manner; we call this *forest carbon accounting* (FCA).

The science of measuring, estimating, and projecting forest carbon stocks and fluxes has become critical to economic and policy decisions, and much of this science employs models. Examples of recent modeling efforts related to FCA include:

* Corresponding author. Tel.: +1 540 231 7674;
fax: +1 540 231 3698.
E-mail address: prisley@vt.edu (S.P. Prisley).

0378-1127/$ – see front matter
doi:10.1016/j.foreco.2004.03.038

Evaluation of ecological and economic impacts of climate and environmental change (Karjalainen, 1996; Mäkipää et al., 1998; Woodbury et al., 1998; Bergh et al., 2003).
Studies of ecological processes and carbon dynamics (Powlson, 1996; Homann et al., 2000; Seely et al., 2002).
Estimation of regional or national forest carbon stocks and potential sequestration (Kurz et al., 1992; Banfield et al., 2002; Joosten et al., 2003; Laclau, 2003; Losi et al., 2003).
Analysis of forest management impacts on terrestrial carbon storage (Jiang et al., 2002; Seely et al., 2002; Díaz-Balteiro and Romero, 2003; Paul et al., 2003; Ritson and Sochacki, 2003).

As negotiations determine the terms under which forest carbon will be accounted, reported, and traded, guidelines and standards are being developed to ensure consistency, accuracy, transparency and verifiability. To date, these guidelines have focused on definitions, data, and reporting—not models. However, estimation of forest carbon is rarely done directly from observational data; a myriad of models are used to estimate forest carbon quantities from surrogate measurements (Kurz et al., 1996; Amthor et al., 2001; Birdsey and Heath, 2001; Smith and Heath, 2002; Jenkins et al., 2003; Smith et al., 2003). For example, modeled biomass expansion factors are used to convert merchantable timber volumes into carbon quantities, and to estimate belowground biomass from aboveground biomass. Carbon in forest pools such as dead wood or litter may be estimated from forest inventory data. Carbon in soils may be modeled based on climate, physiographic region, geographic location, disturbance history, stand age, and other variables. The widespread application of models in forest carbon estimation and accounting suggests that attention to developing guidelines for evaluating and using models is warranted.

2. Purpose

The goal of this paper is to synthesize literature that may inform the development of guidelines for the application of models in policy contexts such as forest carbon accounting. This will be done by reviewing a broad literature on evaluation of models in environmental and ecological applications to find areas of consensus. Then, a review of legal and legislative perspectives on models will identify critical aspects of model evaluation for policy applications. It is far beyond the scope of this paper to evaluate individual forest carbon models or classes of models; an example of such work is provided by Homann et al. (2000). Furthermore, because FCA models are still relatively recent, published modeling guidelines specific to this application do not exist. We draw on the broader literature to find common themes about guidelines for models that can then be applied to FCA models, among others.

For the purposes of this paper, models are defined as abstractions of reality, usually represented by systems of mathematical equations implemented in computer programs, to describe and predict natural phenomena. A similar definition from Helms (1998) defines a model as ‘‘an abstract representation of objects and events from the real world for the purpose of simulating a process, predicting an outcome, or characterizing a phenomenon’’. These three purposes may be viewed as describing, predicting, and estimating. In this paper, we are most interested in the latter, such as models applied in FCA in which the quantitative estimates resulting from the model are paramount.

It is commonly acknowledged that models are essential in many areas of ecological study and environmental management and regulation (Gentil and Blake, 1981; Office of Technology Assessment, 1982; Johnson et al., 1985; Oreskes et al., 1994; Beck et al., 1997; Oreskes, 1998; Landsberg, 2001). The natural systems involved are complex and variable, and defy precise, deterministic quantification (Luis and McLaughlin, 1992; Korzukhin et al., 1996; Olesen, 2001; Parker et al., 2002). When such models are applied to inform public policy, some form of evaluation of a model's reliability is necessary. Oreskes (1998) states ‘‘the demands of good science and the demands of democracy require evidence that the model is reliable’’. Beck et al. (1997) concur, noting ‘‘there is a profound concern, therefore, with the need to establish the validity of a given model in performing a specified task, usually of making predictions of future behavior’’. In a review of early US water resources models in federal and state governments, the Congressional Office of Technology Assessment

(Office of Technology Assessment, 1982) commented that ‘‘lack of accountability has resulted in models that often fail to address decision-makers’ needs for information, require impractical amounts of data, or are not well enough explained to enable others to use them’’. Rykiel (1996) discusses whether models ‘‘require’’ validation at all, and argues that it entirely depends on the purpose for which the model is intended. Others have argued that undue emphasis on validation bears an opportunity cost. Konikow and Bredehoft (1992) suggest that ‘‘The effort spent on model validation would be better spent on developing a more complete understanding of the particular ... problem of interest’’. The preponderance of authors agree, however, that especially in regards to models destined for policy applications, a process for evaluating model suitability is essential.

In this paper, we review literature relevant to the development of guidelines for using models in policy applications. In this process, we will address several questions, such as

1. How are concepts such as validation and verification applied to ecological models?
2. What techniques for model validation/verification have been accepted in the process of scientific peer review of manuscripts about modeling efforts?
3. How have US government agencies with environmental regulatory powers addressed the challenge of verifying/validating ecological models?
4. How have courts in the United States viewed the credibility of models applied in the environmental regulatory or management arenas?

Answers to these questions could support the judicious use of models in areas such as forest carbon accounting, which are critical to policy development and application.

3. Validation, verification, or evaluation

This discussion must begin by addressing terms such as validation and verification. These concepts have been discussed extensively in scientific modeling literature (Loehle, 1997). A variety of definitions are presented in Table 1. Several sources (Schlesinger et al., 1979; Office of Technology Assessment, 1982; ASTM, 1992; Rykiel, 1996) draw a distinction between validation and verification; validation referring to the process of confirming that a conceptual model reflects reality, and verification referring to the process of confirming that computer code accurately reflects the conceptual model. Homann et al. (2000) consider verification to be ‘‘the process of comparing model output with the data that was used to calibrate the model’’ and validation to be comparing model output with independent data not used in calibration. In many other sources, these terms are used interchangeably.

Several authors have made pointed arguments that models cannot be validated; therefore the use of the term itself is misleading (Konikow and Bredehoft, 1992; Oreskes et al., 1994; Oreskes, 1998). This argument is usually related to the concept of hypothesis testing, in which hypotheses can be falsified (rejected) but not accepted. Similarly, the argument goes, it should be possible to falsify or reject models, but they can never be authoritatively validated. Soares et al. (1995) employ the term corroboration, noting that corroboration is when we fail on several attempts to falsify a model.

For these reasons, many authors have adopted the term ‘‘evaluation’’ in place of validation and verification (Soares et al., 1995; Loehle, 1997; Vanclay and Skovsgaard, 1997; Oreskes, 1998; Olesen, 2001; Monserud, 2003; Stage, 2003). We will likewise adopt the terminology of *model evaluation* to encompass the various aspects of confirming the usefulness and reliability of a model.

4. The context for model evaluation

There is no universally appropriate approach to model evaluation. A model must be evaluated in the context of its purpose, domain, and structure (Caswell, 1976; Soares et al., 1995; Beck et al., 1997; Monserud, 2003). Several of the definitions of validation listed in Table 1 include the notion of a specified application (Schlesinger et al., 1979; Brown and Kulasiri, 1996). Beck et al. (1997) summarize Caswell’s (1976) contribution as concluding that ‘‘a judgment about the validity of a model cannot be made in the absence of a specified purpose for the model’’. More recently, Olesen (2001) acknowledges that ‘‘the appropriate

Table 1
Published definitions of model validation, verification, and calibration

Term	Definition	Source
Validation	‘‘Substantiation that a computerized model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model.’’	Schlesinger et al. (1979)
	‘‘Validation is the process of determining how accurately the model can predict real-world events under conditions *different* from those on which the model is developed and calibrated.’’	Office of Technology Assessment (1982)
	‘‘A validated model is one that ’provides a good representation of the actual processes occurring in a real system.’’	International Atomic Energy Agency (1982), cited by Oreskes (1998)
	‘‘The determination that a model indeed reflects the behavior of the real world.’’	US DOE (1986), cited by Oreskes (1998)
	‘‘The process of obtaining assurance that a model, as embodied in a computer code, is a correct representation of the process or system for which it is intended.’’	Davis (1991), cited by Konikow and Bredehoft (1992)
	‘‘A test of the model with known input and output information that is used to assess that the calibration parameters are accurate without further change.’’	ASTM (1992)
	‘‘The process of substantiating that the behavior of the model represents that of the problem entity to satisfactory levels of confidence and accuracy consistent with the intended application of the model within its application domain.’’	Brown and Kulasiri (1996)
	‘‘Validation means that a model is acceptable for its intended use because it meets specified performance requirements.’’	Rykiel (1996)
	‘‘Validation is the determination of the correctness of the model with respect to the user’s needs and requirements.’’	NAPAP (1990), cited by Beck et al. (1997)
	‘‘Validation requires comparing the model predictions with information other than that used in estimating the model. This step is typically an interactive process linked to model calibration.’’	Barton-Aschman Associates (1997)
	‘‘In forest growth modeling, verification and *validation* usually denote qualitative and *quantitative* tests of the model, respectively’’ [emphasis added]	Vanclay and Skovsgaard (1997)
	‘‘the process by which scientists attempt to demonstrate the reliability of a computer model’’	Oreskes (1998)
	‘‘The testing of a model by comparing model results with observations not used to develop the model.’’	Helms (1998)
	‘‘Using subjective opinions regarding the surface, or initial, impression of the model’s realism.’’	Versar Inc. (1988), cited in Beck et al. (1997)
	‘‘. . . the process of comparing model output with the data that was used to calibrate the model. . .’’	Homann et al. (2000)
Verification	‘‘A model is said to be verified when it is determined that the designer’s conception of the model is accurately embodied in the program written and run on the computer.’’	Office of Technology Assessment (1982)
	‘‘Examination of the numerical technique in the computer code to ascertain that it truly represents the conceptual model and that there are no inherent numerical problems with obtaining a solution.’’	ASTM (1992)
	‘‘To say that a model has been verified is to say that its truth has been demonstrated, which implies its reliability as a basis for decision-making’’	Oreskes et al. (1994)
	‘‘Verification is a demonstration that the modeling formalism is correct’’	Rykiel (1996)
	‘‘In forest growth modeling, *verification* and validation usually denote *qualitative* and quantitative tests of the model, respectively’’ [emphasis added]	Vanclay and Skovsgaard (1997)
	‘‘The greater the number and diversity of confirming observations, the more probable it is that the conceptualization embodied in the model is not flawed.’’	Oreskes et al. (1994)
	‘‘. . . comparing model output with data that was not used for calibration.’’	Homann et al. (2000)
Calibration	‘‘A model is considered to be calibrated when model results match experimental observations taken from the particular system under investigation.’’	Office of Technology Assessment (1982)
	‘‘The model is considered calibrated when it reproduces historical data within some subjectively acceptable level of coherence’’.	Konikow and Bredehoft (1992)
	‘‘Calibration is the estimation and adjustment of model parameters and constants to improve the agreement between model output and a data set’’.	Rykiel (1996)

evaluation method depends on the context of the application and the data sets available''. A corollary to this argument is that a model that has been evaluated as appropriate in one context may be utterly inappropriate in another.

In addition to the specified purpose of a model, the type of modeling approach will affect the appropriate method of evaluation. For example, Sharpe (1990) views the concept of model dimensions as a triangle with axes of reality, generality, and precision. Different approaches to model development tend to combine pairs of these dimensions: biostatistical models focus on reality and precision, biomathematical models focus on generality and precision, and forest dynamics models address reality and generality. Sharpe views process modeling as an engineering approach that attempts to address all three. He suggests that each methodology has a distinct criterion for model evaluation: goodness-of-fit for statistics, rigor for mathematics, refutation for science, and performance for engineering.

Battaglia and Sands (1998) discuss the evaluation of process models in forestry, and note that models can be evaluated along the dimensions of resolution (spatial and temporal scale), complexity (number of environmental variables and processes included), and generality (situations and systems to which the model can be applied). Usually, in the process of model design and development, increases in one dimension (i.e., higher resolution, increased complexity, or broad applicability) must be offset by decreases in another. Similarly, there is often a tradeoff between an increase in explanatory power and generality. A model might perform very well in a narrow scope of circumstances, or it might perform at a barely acceptable level across a broad range of conditions. Because model developers often balance these outcomes (Håkanson, 1995), model evaluators should consider them as well. A model designed to be broadly applied is unlikely to perform as precisely in a given application or geographic region as a model designed exclusively for that application/domain.

5. Components of model evaluation

Beck et al. (1997) refer to two aspects of a model that are relevant to evaluation: *composition* and *performance*. *Composition* of a model refers to the manner in which constituent hypotheses are formulated and assembled. *Performance* refers to the acceptability and usefulness of model outputs for an intended task. While composition is an internal measure of model reliability, performance is an external measure. These two aspects of model evaluation are echoed repeatedly in model evaluation literature, and each has its own logical means of assessment.

Loehle (1997) suggests that the *composition* criterion (as indicated by biological and ecological realism) outweighs the importance of *performance*: ''It is not sufficient that a model fits field data if it does so by employing biologically unreasonable behaviors or processes''. Johnson et al. (1985) illustrate this point with a critique of a model of waterfowl populations. The model they discuss provided an adequate goodness-of-fit with data, yet contained biologically untenable premises that would mislead decision-makers when applied to situations only slightly different than those under which the model was developed. Furthermore, it should be noted that evaluation of *composition* goes beyond the judging the validity of constituent hypotheses (components); the connections and interrelationships of these components must reflect ecological reality. As Loehle (1997) notes, ''It is at the ecosystem level that we observe the integrated effect of the multiple processes at work in a model... Ecosystem-level test criteria thus strengthen tests of integrated model behaviors, though they do not verify that each model component is correct.''

The *composition/performance* view of models gives rise to two approaches to model evaluation: (1) scientific peer review, and (2) statistical comparisons of model results with field observations (ASTM, 2000; Olesen, 2001). Logically, scientific peer review can address the *composition* of a model, while statistical comparisons with field data can address model *performance*. Comparison of model results with observational data is frequently termed ''history matching'' in the literature. It should be noted that many definitions of model validation consider only history matching; they do not address a model composition component (Schlesinger et al., 1979; ASTM, 1992; Vanclay and Skovsgaard, 1997; Helms, 1998).

Parker et al. (2002) view these two components as anachronistic and insufficient, stating ''At one time, 'history matching' and 'peer review' were the two necessary and sufficient cornerstones of the [model validation] process. But...[there] has come a dissatisfaction with the sufficiency of these conventional cornerstones''. They present three questions that must be answered about a model: (1) Has the model been constructed of approved constituent hypotheses? (2) Does its behavior approximate that observed in reality? (3) Does it fulfill its designated task or purpose? They acknowledge the roles of peer review and history matching as addressing the first two of these questions, but note a lack of developed tools or approaches for addressing the third.

A third approach to model evaluation that appears with regularity in the literature is sensitivity analysis (Newberry and Stage, 1988; ASTM, 1992; Soares et al., 1995; Barton-Aschman Associates, 1997; Loehle, 1997; Beck and Chen, 2000; Potter et al., 2001; Losi et al., 2003; Paul et al., 2003). Sensitivity analysis addresses issues such as model robustness, applicability beyond the range of data used to fit the model, stability of model parameters, and variability of model outputs.

To summarize, a reader can find numerous recommendations in the literature for essential components to model evaluation: Gentil and Blake (1981) use a flow chart illustration, ASTM (1992) describes seven components, Soares et al. (1995) list five, Vanclay and Skovsgaard (1997) summarize using five. In most cases, these components of model evaluation can be synthesized into three: (1) scientific peer review of model composition, (2) quantitative analysis of model results compared to field observations, and (3) sensitivity analysis.

6. Peer review

Beck et al. (1994) describe a variety of challenges in evaluating model composition through the peer review process. As modeling becomes more commonplace, it is increasingly difficult to get descriptions of model development and design published, and therefore subjected to the peer review process of scientific journals. Furthermore, the peer review of a manuscript describing model development and application is notably different than peer review of the model itself. A rigorous review of a model might include running the model against independent data sets, or comparing the outputs from one model to those of a similar model with the same inputs. An example of such an effort can be found in the VEMAP project (Schimel et al., 1997, 2000), in which three ecosystem models were compared using the same input data sets and parameters. Results of the VEMAP project provided insights into the spatial and temporal variability of terrestrial carbon storage as estimated by the three models. Such comprehensive model comparisons are expensive, ''expert-intensive'', time-consuming, and consequently rare.

Parker et al. (2002) note the challenge of forming an adequate peer review panel for ecological models:

> ''The constitution of a 'peer group' can therefore be expected to be very different and more varied than just the former sub-groups of model builders and model users, steeped largely in the professional training and standards of science. While there may be scope for modest parts of evaluation by mono-disciplinary (scientific) peers, there will be few renaissance (wo)men capable of reviewing the whole and fewer still who will be able to claim no conflict of interest as the model evolves over possibly many years in the light of successive reviews by these few''.

Beck et al. (1994) also emphasize that ''as models have become increasingly complex and aggregate science across more than one discipline, it is increasingly clear that more than one subject matter expert is required to provide adequate scientific review''. Forest carbon accounting models are an excellent example of such complexity. For instance, the development of the carbon budget model for the Canadian forest sector (Kurz et al., 1992) required experts from a wide array of disciplines with knowledge of the specific forest ecosystems being modeled. The scope of such a model would require a comparably large array of experts for a thorough independent peer review.

The US EPA has issued guidance noting that peer review should be tied to temporal steps in the process of model design, development, application, and revision (US EPA, 1994a). When models persist in use over decades, it becomes important to subject them to repeat reviews as the science underlying the constituent hypotheses matures and as observations of different conditions become available. In addition, when a model is applied in a novel application, a different geography, or for an unforeseen purpose, additional peer review may be called for.

Beck et al. (1994) propose a three-part strategy for peer review of models: (1) reference code, version documentation, and test data set maintenance, (2) publication in refereed journals, and (3) periodic or issue-specific group peer reviews. To facilitate open scientific review of models, ASTM (1992) recommends a list of required components of model documentation and suggests that, at a minimum, models be described in scientific literature and that a user's manual be available before significant application.

7. Quantitative analysis of model results

The quantitative comparison of model results against observations is perhaps the most familiar aspect of model validation. This topic has been discussed widely in the literature, with near-universal recognition of the value of this process. However, as noted by Mäkipää et al. (1998), "…there have been comparatively few quantitative evaluations of model projections to the long-term observations." Many authors acknowledge that substantial difficulties in such comparisons remain. Three primary obstacles include: (1) availability of adequate independent data sets for validation, (2) selection of meaningful statistical tests or approaches, and (3) specification of what constitutes acceptable performance.

One of the foremost challenges to quantitative evaluation of model performance is the availability of data of useful spatial and temporal resolution and measurement precision. For example, Monserud (2003) notes that forest gap models are rarely validated because of a lack of suitable validation data. When such data are lacking, he notes, "reasonable model behavior is often determined by expert opinion". Numerous authors and definitions explicitly state that validation should be conducted with *independent* data sets not involved in model construction or calibration (OTA, 1982; Barton-Aschman Associates, 1997; Helms, 1998). However, Mäkelä et al. (2000) argue that this requirement "presumes the existence of 'independent' data, ignores the expense associated with its collection, and misdirects attention away from the more critical needs for a rich set of conditions represented in evaluation data".

A wide variety of statistical techniques have been applied to the evaluation of model performance as defined by comparing model results to observed data. These include graphical or visualization comparisons (Rykiel, 1996; Butler et al., 1999), paired statistical comparison metrics (ASTM, 2000), hypothesis testing (Luis and McLaughlin, 1992), significance levels or confidence limits for model parameters (Chen et al., 1998), regression analysis (US EPA, 1994b), correlation computation (Brown and Kulasiri, 1996), bootstrapping (Butler et al., 1999), goodness-of-fit tests, simulation, and many others.

A third challenge in model performance evaluation is specification of the criteria that will need to be met to judge a model as "acceptable". In fact, there is disparity in viewpoints about whether model validation is a binary decision (a model is acceptable or it is not), or a whether it is a measure of degree (Beck et al., 1997). Newberry and Stage (1988) describe a validation procedure with four possible outcomes, including "model is adequate", "model needs revision", "data are inadequate to evaluate model", and "model is irrelevant". In ASTM (1992) standards, varying levels of acceptable accuracy are tied to model purposes. For example, when models are used for "range finding", an order of magnitude might be acceptable accuracy; a model employed for prediction might require an accuracy of 5 to 15%.

It should also be noted that the purpose of the model may make specification of model acceptability requirements simple or complicated. For example, a model may be used to assess the risk that children will have lead concentrations in their blood that exceed a prescribed limit (e.g., US EPA, 1994b). In this case, the specification for model acceptability may be quite straightforward, and expressed in terms of confidence limits on the predicted lead concentration. Other models, such as forest ecosystem process models, produce a series of projected conditions over time (Mäkelä et al., 2000). In this case, model evaluation must consider the trajectory of numerous variables over time, and specification of a simple rule for model acceptance is quite challenging. Forest carbon accounting models are more similar to the latter example, in that estimates of change in numerous interacting carbon pools are needed over time. It is evident from literature that the a priori specification of decision rules continues to challenge ecological modelers. Rykiel (1996) suggests that "[T]he most common problem with ecological and environmental models is failure to state what the validation criteria are.''

So while quantitative analysis of model performance is widely accepted as a critical part of model evaluation, there is no universally applicable approach to either the statistical tests or the framing of acceptance criteria, and availability of validation data is often cited as a primary impediment.

8. Sensitivity analysis

Sensitivity analysis examines the degree to which the model result is affected by changes in selected input parameters (ASTM, 1992). Because ecological and environmental models are frequently (and necessarily) applied outside the range of conditions represented by the data used to create them, sensitivity analysis is especially important (Beck et al., 1997). Sensitivity analysis provides insight into the influence of different model parameters, which can promote a general understanding of model robustness. When a model is shown to be relatively insensitive to a given parameter, then relatively less time and effort can be expended on selecting or estimating the parameter value, and confidence in the stability of model results may increase (Chen and Beck, 1999). An example of such an analysis in a forest carbon accounting context is described by Smith and Heath (2001), who identified model sensitivity to range, shape, and covariability of model parameters. Likewise, the VEMAP project described previously conducted simultaneous sensitivity analysis with multiple models using the same range of inputs.

Despite its strengths, sensitivity analysis cannot be viewed in the same light as model validation. Loehle (1997) argues that simply because an analysis shows a model to be insensitive to modest parameter error, the model cannot be deemed ''validated''. Similarly, when model outputs vary widely in response to perturbations in parameter values, the model is not necessarily invalidated (Beck et al., 1997).

9. Existing guidelines and standards

Some of the recommendations reviewed herein have made it into standards and guidelines for model evaluation and application. For example, the American Society for Testing and Materials (ASTM) has standards for statistical evaluation of atmospheric dispersion model performance (ASTM, 2000), and a guide for comparing groundwater flow model simulations to site-specific information (ASTM, 2002). A standard practice for evaluating mathematical models for the environmental fate of chemicals was adopted (ASTM, 1992) but subsequently discontinued.

The US Department of Transportation has published a ''Model Validation and Reasonableness Checking Manual'' (Barton-Aschman Associates, 1997) for the evaluation of travel models. The US Environmental Protection Agency (EPA) has expended considerable effort at formalizing guidelines for model evaluation. Results include guidelines for peer review of models (US EPA, 1994a), a white paper on issues related to adopting model use acceptability guidelines (US EPA, 1999), and guidelines for developing quality assurance plans for modeling efforts (US EPA, 2002).

In Europe, over a decade of effort has gone into the harmonization of atmospheric dispersion models (Olesen, 2001). A significant result of this collaboration has been the development of a standardized approach and tool set to compare models using common data: the ''Model Validation Kit''.

10. Policy and legal concerns for the use of models

The use of scientific models to inform agency decisions, to promulgate regulations, or to otherwise initiate agency actions will invariably imply certain legal and policy constraints. Those constraints are often the result of the reliance upon legalistic mechanisms to contest or force agency action (Kagan, 2001). As such, the proposed use of scientific models must consider the potential for judicial scrutiny of the model, the model's particular application, as well as the policies followed in developing and adopting the model. These considerations are particularly important when use of a predictive model will either facilitate or restrict action by the agency or by agency-regulated private parties. Such predictive models are typically an amalgam of political, regulatory, and scientific components (Westerbrook, 1999), the outputs of which are often the source of, or a contributor to, legal disputes.

Federal courts in the United States have a lengthy history of involvement in arbitrating the use of predictive models by federal agencies (Case, 1982). Should a court invalidate the application of a particular model, an agency's ability to use the model may be impaired or even eliminated. Resolving those disputes may well entail a judicially determined mixture of science, policy, and social concerns (Fienberg et al., 1995). As such, it has been suggested that agency policy considerations should be explicit, particularly when an agency chooses one set of assumptions over another, draws one set of inferences over another, or chooses a particular predictive model over another (Administrative Conference of the US, 1985; Crawford-Brown, 2001).

11. Data Quality Act

Of particular future significance for US modelers engaged in the development of agency policy, regulations, or other proposed actions will be the as-yet-untested effects of the Data Quality Act (Pub. L. No. 106-554, 114 Stat. 2763A-153 [2000]). This legislation requires that nearly all federal agencies prepare guidance to maximize the quality, objectivity, utility, and integrity of information the agency publicizes or disseminates—including, presumably, quantitative models. It further provides for a procedural mechanism for aggrieved parties to request that the agency correct information perceived to violate the guidance prepared by the agency or the Office of Management and Budget. In response to this legislation, the Department of Agriculture promulgated guidance covering the dissemination of both regulatory information and scientific research information (USDA, 2003a,b). This guidance includes the following:

When creating estimates or forecasts that are derived from existing data sources *using models* or other techniques [emphasis added]:

> Use sound statistical methods that conform to accepted professional standards.
> Document models and other estimation or forecasting techniques to describe the data sources used and the methodologies and assumptions employed.

In addition, the USDA guidelines cite the Office of Management and Budget (OMB) recommendations for conducting scientific peer review, which should be applied to such models:

OMB recommends that

(a) peer reviewers be selected primarily on the basis of necessary technical expertise,
(b) peer reviewers be expected to disclose to agencies prior technical/policy positions they may have taken on the issues at hand,
(c) peer reviewers be expected to disclose to agencies their sources of personal and institutional funding (private or public sector), and
(d) peer reviews be conducted in an open and rigorous manner.

The regulatory guidelines under the Data Quality Act are far-ranging, including such things as rulemaking, documents prepared in accordance with requirements of the National Forest Management Act (including national forest plans), environmental documents prepared under the National Environmental Policy Act, and biological evaluations and assessments prepared under the Endangered Species Act. The USDA guidance emphasizes:

> use of sound analytical methods;
> use of reasonably reliable and timely data;
> ensuring transparency of analysis and documentation of data sources, including uncertainty and limitations;
> where appropriate, employing external peer review.

These criteria largely mirror the various criteria identified to maximize the survivability of predictive models in regulatory and judicial settings.

To date, the Forest Service reports 11 Data Quality Act requests, eight for informational corrections and three for decision reconsideration (USDA, 2003b). None of the 11 has questioned the use of a predictive model. The potential for challenges to such models, however, should not be overlooked. Close adherence to the existing agency guidance, and amendment of the guidance as necessary to reflect evolving judicial concerns should be considered. Though the general expectation is that a court will uphold an agency's use of a predictive model (Otero-Phillips, 1998), failure to adhere to internal agency guidelines, or the guidance established in legal precedence will increase the chance that the use of a particular model will be curtailed.

12. Legal views on model verification

While there are decided parallels between the accepted scientific practices in evaluating models and the legal considerations for model use, the differing standards applied by scientists and courts of law in verifying a predictive model present unique concerns for public bureaucracies (Grossman, 1992). An example is the role that the opportunity for public comment plays in the judicial calculus. While certainly not a matter of scientific concern, it has been suggested that failure to afford a meaningful comment period on a model, wherein objections to the model might be raised, should serve as a fatal legal defect in the future use of the model (Case, 1982). Additionally, at least one court ratified the use of a Forest Service predictive model, in part due to the agency providing for and considering public comments (Sierra Club v. US Forest Service, 1993).

Various sources have identified criteria that may or should be considered in assessing the propriety of a particular model's application. They include: stating the model's assumptions (Case, 1982; Administrative Conference of the US, 1985; Shook and Tartal, 2000); disclosing limitations (Administrative Conference of the US, 1985; Crawford-Brown, 2001); disclosing the decision-making variables not dependent on the model (Administrative Conference of the US, 1985); disclosing uncertainty (Administrative Conference of the US, 1985; Shook and Tartal, 2000); the degree of scientific acceptability (Grossman, 1992; Shook and Tartal, 2000; Crawford-Brown, 2001); the use of peer review (Shook and Tartal, 2000), and adequate empirical testing (Jones, 1987; Grossman, 1992; Grossman and Gagne, 1993; Sklash et al., 1999).

13. Appropriate model application

Of fundamental importance is the realization that predictive models relied upon for uses other than those for which they were developed may subsequently result in poorly designed polices (House, 1982). The process or manner in which a model is applied can be of equal or even greater importance than the substance of the model (Shook and Tartal, 2000). Misuse or erroneous application may lead to rejection of the model's results. Additionally, at least one court has noted that an overly rigid application of the model by the agency will subject the model and the agency's supporting evidence to heightened judicial scrutiny (Chemical Manufacturers Association v. EPA, 1994).

14. Evaluation of model results

Discrepancies between model predictions and actual data obtained by monitoring or other means have certainly not gone unnoticed or unchallenged by various courts. Unfortunately, there is no consensus on the question of how inconsistencies with observed data will affect the legal acceptability of a particular model (Johnson, 1987). Additionally, of particular difficulty for federal agencies in developing predictive models and the protocols for their use is the fractured nature of the US federal judiciary system. Due to the number of federal district courts and appellate circuits, diverse opinions may develop on the creation and use of predictive models (Johnson, 1987). In such instances, agencies may receive confusing, or even conflicting messages as to the acceptability of the subject model.

15. Example cases

There is but a single reported case detailing Forest Service reliance on a predictive model (Sierra Club v. USFS, 1993). In that case, the court determined that the agency's use of the HABCAP program was legitimate provided the Forest Service revealed the data and assumptions of the model, allowed and considered public comment on the model, and ensured that the ultimate decision rested with the agency, not with the model output.

In contrast, the Environmental Protection Agency's use of its CRSTER model was determined to be arbitrary and capricious (Ohio v. EPA, 1986). The court noted specifically that the agency failed to calibrate the model with site-specific data, and by failing to do so, the EPA violated its own modeling guidelines. While the facts of the cases differed, it is important to note the emphasis placed upon procedures by both courts. While a court may be loath to second-guess the agency's choice of a model, or critique the model's inner workings, courts are much more receptive to ensuring that appropriate processes are followed prior to the employment of the model, particularly transparency in the model's assumptions and availability of the underlying data.

16. Consensus recommendations and summary

There is abundant literature on the topics of model evaluation approaches, guidelines for application of models in policy settings, and standards for model documentation. If models are to be widely applied in the context of reporting carbon stores and fluxes for greenhouse gas accounting (or for carbon markets), it is reasonable to expect that these models should adhere to scientifically relevant and judicially proven guidelines. The following recommendations are made repeatedly in the literature, and seem relevant in the context of forest carbon accounting.

1. The scope of the model should be clearly defined. This is the model domain, and can be expressed in terms of ecophysiographic regions, spatial scale, and temporal scale. For FCA models, the domain would also include the range of silvicultural treatments that are modeled, tree species, forest product end uses, etc. The model application should be limited to the domain for which a model has been developed and evaluated. Models should be applied to additional situations only after appropriate validation.
2. Models should be clearly documented. Documentation should include assumptions, known limitations, embedded hypotheses, assessment of uncertainties, and sources (for equations, data sets, factors or parameters, etc.). Thorough documentation of model processes and assumptions should allow independent investigators to reproduce model results from the same input data.
3. Models should be scientifically reviewed. A thorough peer review process would include evaluation of equations, modeling system, software, and calibration data set, for applicability and adequacy. In addition and as appropriate, the review should be conducted not only by modeling specialists, but by experts in relevant fields of biology, ecology, physiology, etc. This presents a particular challenge for FCA models, which include components spanning a wide range of technical disciplines and potentially broad array of ecophysiographic regions.
4. When possible, model results should be compared with field observations and results of this comparison should be published. Such model comparisons are most valuable when done in the domain of model application. This may be more feasible for some components of FCA models than others; for example, vegetation biomass is far more readily validated than soil organic carbon.
5. Sensitivity analysis should be conducted to identify behavior of model across the range of parameters for which it is to be applied. This recommendation is perhaps the most readily applied to FCA models, as evidenced by examples (Schimel et al., 2000; Smith and Heath, 2001).
6. Models should be made available for testing/evaluation. For example, an online implementation of a forest carbon estimator is being developed which will allow quantification of forest carbon storage for user-specified areas within the US (Proctor et al., 2004).
7. Because models are a function of the scientific understanding and data at the point in time at which the model was developed, they should be periodically reviewed in light of new knowledge and data. If necessary, models should be recalibrated based on this evaluation. In an area of widespread scientific inquiry and extensive research such as the forest carbon cycle, new knowledge may indicate that FCA models should be reviewed on decadal time frame, if not more frequently.
8. When models are applied in regulatory or policy development, a public comment period is critical. While forest carbon accounting is unlikely to attract widespread public interest, interested parties include forest managers, landowners, forest products buyers, and scientists. If and when markets for carbon trading are more firmly established, more parties will become financially involved and interested in mechanics and assumptions of FCA models.

It is also clear that consistency and openness in the process of developing models, and a well-defined and appropriate context for applying models, are crucial in providing a model application that will withstand public scrutiny and legal challenge.

When ecological or environmental models are applied in settings with significant policy, economic, regulatory, or social impacts, it is reasonable to hold them to high standards. When FCA models provide information that is subsequently used in the context of international agreements, policy development, regulation, or emissions trading markets, it seems clear that similar high standards should apply. Substantial scientific discussion and judicial action have addressed model development, evaluation, and application in environmental regulatory arenas. In some cases (within some agencies and for some specific applications), standards and guidelines have been developed. It appears that the application of forest carbon accounting models would benefit from broader discussion and consideration of similar standards and guidelines.

Acknowledgements

Funding for this effort was provided by the USDA Forest Service Northeastern Research Station, Project NE-4104. The authors gratefully acknowledge Dr. Linda S. Heath, Dr. Marilyn A. Buford, and Dr. Linda Joyce of the USDA Forest Service, Dr. Matt Fladeland of NASA, and several anonymous reviewers for helpful comments and additional sources.

References

Administrative Conference of the United States, 1985. Agency procedures for performing regulatory analysis or rules (Recommendation No. 85-2). 1 Code of Federal Regulations § 305.85-2, Washington, DC.

American Society for Testing and Materials (ASTM), 1992. Standard Practice for Evaluating Mathematical Models for the Environmental Fate of Chemicals. Standard E978-92. ASTM, West Conshohocken, Pennsylvania.

American Society for Testing and Materials (ASTM), 2000. Standard Guide for Statistical Evaluation of Atmospheric Dispersion Model Performance. Standard D6589-00. ASTM, West Conshohocken, Pennsylvania.

American Society for Testing and Materials (ASTM), 2002. Standard Guide for Comparing Ground-Water Flow Model Simulations to Site-Specific Information. Standard D5490-93. ASTM, West Conshohocken, Pennsylvania.

Amthor, J.S., Chen, J., Clein, J., Frolking, S., Goulden, M., Grant, R., Kimball, J., King, A., McGuire, A., Nikolov, N., Potter, C., Wang, S., Wofsy, S., 2001. Boreal forest CO_2 exchange and evapotranspiration predicted by nine ecosystem process models: inter-model comparisons and relationships to field measurements. J. Geophys. Res. 106 (D24), 33, 623–633 and 648.

Banfield, G.E., Bhatti, J.S., Jiang, H., Apps, M.J., 2002. Variability in regional scale estimates of carbon stocks in boreal forest ecosystems: results from West-Central Alberta. For. Ecol. Manage. 169, 15–27.

Barton-Aschman Associates, Inc., 1997. Model Validation and Reasonableness Checking Manual. Travel model improvement program. Federal Highway Administration. http://tmip.fhwa.dot.gov/clearinghouse/docs/mvrcm/ (accessed 13 May 2003).

Battaglia, M., Sands, P.J., 1998. Process-based forest productivity models and their application in forest management. For. Ecol. Manage. 102, 13–32.

Beck, M.B., Mulkey, L.A., Barnwell, T.O., 1994. Model Validation for Predictive Exposure Assessments. Report to EPA, 52 pp. http://cfpub.epa.gov/crem/library/whitepaper_1994.pdf.

Beck, M.B., Ravetz, J.R., Mulkey, L.A., Barnwell, T.O., 1997. On the problem of model validation for predictive exposure assessments. Stoch. Hydrol. Hydraul. 11, 229–254.

Beck, M.B., Chen, J., 2000. Assuring the quality of models designed for predictive tasks. Sensitivity Analysis. Wiley, pp. 401–420.

Bergh, J., Freeman, M., Sigurdsson, B., Kellomäki, S., Laitinen, K., Niinistö, S., Peltola, H., Linder, S., 2003. Modelling the short-term effects of climate change on the productivity of selected tree species in Nordic countries. For. Ecol. Manage. 183, 327–340.

Birdsey, R.A., Heath, L.S., 2001. Forest inventory data, models, and assumptions for monitoring carbon flux. In: Soil Carbon Sequestration and the Greenhouse Effect. Soil Science Society of America, Madison, WI. SSSA Special Publication no. 57, Chapter 10, pp. 125–135.

Brown, T.N., Kulasiri, D., 1996. Validating models of complex, stochastic, biological systems. Ecol. Model. 86, 129–134.

Butler, B.P., Cox, M.G., Forbes, A.B., Harris, P.M., Lord, G.J., 1999. Model validation in the context of metrology: a survey. NPL Report CISE 19/99. National Physical Laboratory, Teddington, Middlesex, UK, 69 pp.

Case, C.D., 1982. Problems in judicial review arising from the use of computer models and other quantitative methodologies in environmental decisionmaking. Boston College Environ. Affairs Law Rev. 10 (4), 251–363.

Caswell, H., 1976. The validation problem. In: Patten, B.C. (Ed.), Systems Analysis and Simulation in Ecology, vol. IV. Academic Press, New York, pp. 313–325.

Chemical Manufacturers Association v. EPA, 1994. 28 F.3d 289 (D.C. Cir.).

Chen, J., Beck, M.B., 1999. Quality Assurance of Multi-media Model for Predictive Screening Tasks (EPA/600/R-98/106). US EPA, Washington, DC, 63 pp.

Chen, Y.D., McCutcheon, S.C., Norton, D.J., Nutter, W.L., 1998. Stream temperature simulation of forested riparian areas: model application. J. Environ. Eng. 124 (4), 316–328.
Crawford-Brown, D., 2001. Scientific models of human health risk analysis in legal and policy decisions. Law Contemp. Probl. 64 (4), 63–81.
Díaz-Balteiro, L., Romero, C., 2003. Forest management optimisation models when carbon captured is considered: a goal programming approach. For. Ecol. Manage. 174, 447–457.
Economist, 2002. The Chicago Climate Exchange: Trading Hot Air.
Fienberg, S.E., Krislov, S.H., Straf, M.I., 1995. Understanding and evaluating statistical evidence in litigation. Jurimetrics J. 36, 1–28.
Gentil, S., Blake, G., 1981. Validation of complex ecosystem models. Ecol. Model. 14, 21–38.
Grossman, S.J., 1992. Science and scientific evidence. IDEA: The J. Law Technol. 32, 179–202.
Grossman, S.J., Gagne, C.K., 1993. Science and scientific evidence II. Connecticut Law Rev. 25, 1053–1065.
Håkanson, L., 1995. Optimal size of predictive models. Ecol. Modell. 78, 195–204.
Helms, J.A. (Ed.), 1998. The Dictionary of Forestry. Society of American Foresters, Bethesda, MD, 210 pp.
Homann, P.S., McKane, R.B., Sollins, P., 2000. Belowground processes in forest-ecosystem biogeochemical simulation models. For. Ecol. Manage. 138, 3–18.
House, P.W., 1982. The Art of Public Policy Analysis. Sage, Beverly Hills, CA.
Jenkins, J.C., Chojnacky, D.C., Heath, L.S., Birdsey, R.A., 2003. National-scale biomass estimators for United States tree species. For. Sci. 49 (1), 12–34.
Jiang, H., Apps, M.J., Peng, C., Zhang, Y., Liu, J., 2002. Modelling the influence of harvesting on Chinese boreal forest carbon dynamics. For. Ecol. Manage. 169, 65–82.
Johnson, D.H., Conroy, M.J., Nichols, J.D., 1985. The need for accuracy in modeling: an example. Ecol. Model. 30, 157–161.
Johnson, W.B., 1987. Application of air quality modeling to decisionmaking under the Clean Air Act (42 USCA §§7401–7626). Am. Law Rep. 84, 710.
Jones, E.M., 1987. A component approach to minimal rationality review for agency rulemaking. Admin. Law Rev. 39, 275–313.
Joosten, R., Schumacher, J., Wirth, C., Schulte, A., 2003. Evaluating tree carbon predictions for beech (*Fagus sylvatica* L.) in western Germany. For. Ecol. Manage. 189, 87–96.
Kagan, R.A., 2001. Adversarial Legalism: The American Way of Life. Harvard University Press, Cambridge, MA.
Karjalainen, T., 1996. Dynamics and potentials of carbon sequestration in managed stands and wood products in Finland under changing climatic conditions. For. Ecol. Manage. 80, 112–132.
Konikow, L.F., Bredehoft, J.D., 1992. Ground-water models cannot be validated. Ad. Water Resour. 15, 75–83.
Korzukhin, M.D., Ter-Mikaelian, M.T., Wagner, R.G., 1996. Process versus empirical models: which approach for forest ecosystem management? Can. J. For. Res. 26, 879–887.
Kurz, W.A., Apps, M.J., Webb, T.M., McNamee, P.J., 1992. The Carbon Budget of the Canadian Forest Sector: Phase I. Forestry Canada, Edmonton, 93 pp.
Kurz, W.A., Beukema, S.J., Apps, M.J., 1996. Estimation of root biomass and dynamics for the carbon budget model of the Canadian forest sector. Can. J. For. Res. 26 (11), 1973–1979.
Laclau, P., 2003. Biomass and carbon sequestration of ponderosa pine plantations and native cypress forests in northwest Patagonia. For. Ecol. Manage. 180, 317–333.
Landsberg, J., 2001. Modelling forest ecosystems: state-of-the-art, challenges, and future directions. Can. J. For. Res. 33 (3), 385–397.
Loehle, C., 1997. A hypothesis testing framework for evaluating ecosystem model performance. Ecol. Model. 97, 153–165.
Losi, C.J., Siccama, T.G., Condit, R., Morales, J.E., 2003. Analysis of alternative methods for estimating carbon stock in young tropical plantations. For. Ecol. Manage. 184, 355–368.
Luis, S.J., McLaughlin, D., 1992. A stochastic approach to model validation. Adv. Water Resour. 15, 15–32.
Mäkelä, A., Landsberg, J.J., Ek, A.E., Burk, T.E., Ter-Mikaelian, M., Ågren, G., Oliver, C.D., Puttonen, P., 2000. Process-based models for forest ecosystem management: current state-of-the-art and challenges for practical implementation. Tree Physiol. 20, 289–298.
Mäkipää, R., Karjalainen, T., Pussinen, A., Kukkola, M., Kellomäki, S., Mälkönen, E., 1998. Applicability of a forest simulation model for estimating effects of nitrogen deposition on a forest ecosystem: test of the validity of a gap-type model. For. Ecol. Manage. 108, 239–250.
Monserud, R.A., 2003. Evaluating forest models in a sustainable forest management context. For. Biometry Model. Inform. Sci. 1, 35–47. http://www.fbmis.info.
Murray, B.C., Prisley, S.P., Birdsey, R.A., Sampson, R.N., 2000. Carbon sinks in the Kyoto Protocol: potential relevance for US Forests. J. For. 98 (9), 6–11.
Newberry, J.D., Stage, A.R., 1988. Validating forest growth models: procedures defined by resource decisions. In: Ek, A.R., Shifley, S.R., Burk, T.E. (Eds.), Forest Growth Modelling and Prediction: Proceedings of the IUFRO Conference, 23–27 August 1987, Minneapolis, MN. USFS General Technical Report NC-120, pp. 786–793.
Office of Technology Assessment, 1982. Use of Models for Water Resources Management, Planning, and Policy. US OTA, Washington, DC, 242 pp.
Olesen, H.R., 2001. Ten years of harmonization activities: past, present, and future. In: Proceedings of the Seventh International Conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes. http://www.harmo.org/Docs/TenYears.pdf.
Olesen, H.R., 2001. A platform for model validation. In: Proceedings of the Seventh International Conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes. http://www.harmo.org/Docs/ME_platform.pdf.
Oreskes, N., Shrader-Frechette, K., Belitz, K., 1994. Verification, validation, and confirmation of numerical models in the earth sciences. Science 263, 641–646.

Oreskes, N., 1998. Evaluation (not validation) of quantitative models. Environ. Health Perspect. 106 (Suppl. 6), 1453–1460.

Otero-Phillips, C.V., 1998. What's in the forecast? A look at the EPA's use of computer models in emissions trading. Rutgers Comput. Technol. Law J. 24, 187–222.

Parker, P., Letcher, R., Jakeman, A., Beck, M.B., Harris, G., Argent, R.M., Hare, M., Pahl-Wostl, C., Voinov, A., Janssen, M., Sullivan, P., Scoccimarro, M., Friend, A., Sonnenshein, M., Barker, D., Matejicek, L., Odulaja, D., Deadman, P., Lim, K., Larocque, G., Tarikhi, P., Fletcher, C., Put, A., Maxwell, T., Charles, A., Breeze, H., Nakatani, N., Mudgal, S., Naito, W., Osidele, O., Eriksson, I., Kautsky, U., Kautsky, E., Naeslund, B., Kumblad, L., Park, R., Maltagliati, S., Girardin, P., Rizzoli, A., Mauriello, D., Hoch, R., Pelletier, D., Reilly, J., Olafsdottir, R., Bin, S., 2002. Progress in integrated assessment and modeling. Environ. Model. Softw. 17, 209–217.

Paul, K.I., Polglase, P.J., Richards, G.P., 2003. Predicted change in soil carbon following afforestation or reforestation, and analysis of controlling factors by linking a C accounting model (CAMFor) to models of forest growth (3PG), litter decomposition (GENDEC) and soil C turnover (RothC). For. Ecol. Manage. 177, 485–501.

Potter, C.S., Wang, S., Nikolov, N.T., McGuire, A.D., Liu, J., King, A.W., Kimball, J.S., Grant, R.F., Frolking, S.E., Clein, J.S., Chen, J.M., Amthor, J.S., 2001. Comparison of boreal ecosystem model sensitivity to variability in climate and forest site parameters. J. Geophys. Res. 106 (33), 671–688.

Powlson, D.S., 1996. Why evaluate soil organic matter models? In: Powlson, D.S., Smith, P., Smith, Jo.U. (Eds.), Evaluation of Soil Organic Matter Models Using Existing Long-term Datasets. Spinger-Verlag, Berlin. NATO ASI Series, Series I, Global Environmental Change, vol. 38, pp. 3–11.

Proctor, P., Heath, L.S., Van Deusen, P.C., Gove, J.H., Smith, J.E., 2004. COLE: a web-based tool for interfacing with forest inventory data. In: McRoberts, R., Reams, G., Van Deusen, P., McWilliams, W. (Eds.), Proceedings of the Fourth Annual Forest Inventory Analysis Symposium, November 19–21, 2002, New Orleans, LA. General Technical Report NC. US Department of Agriculture, Forest Service, North Central Research Station, St. Paul, MN.

Ritson, P., Sochacki, S., 2003. Measurement and prediction of biomass and carbon content of *Pinus pinaster* trees in farm forestry plantations, south-western Australia. For. Ecol. Manage. 175, 103–117.

Rykiel Jr., E.J., 1996. Testing ecological models: the meaning of validation. Ecol. Model. 90, 229–244.

Schlesinger, S., Crosbie, R.E., Gagné, R.E., Innis, G.S., Lalwani, C.S., Loch, J., Sylvester, R.J., Wright, R.D., Kheir, N., Bartos, D., 1979. Terminology for model credibility. SCS Technical Committee on Model Credibility. Simulation 32 (3), 103–104.

Seely, B., Welham, C., Kimmins, H., 2002. Carbon sequestration in a boreal forest ecosystem: results from the ecosystem simulation model, FORECAST. For. Ecol. Manage. 169, 123–135.

Sharpe, P.J.H., 1990. Forest modeling approaches: compromises between generality and precision. In: Dixon, R.K., Meldahl, R.S., Ruark, G.A., Warren, W.G. (Eds.), Process Modeling of Forest Growth Responses to Environmental Stress. Timber Press, Portland, OR, pp. 180–190.

Schimel, D.S., VEMAP Participants, Braswell, B.H., 1997. Continental scale variability in ecosystem processes: models, data, and the role of disturbance. Ecol. Monogr. 67 (2), 251–271.

Schimel, D., Melillo, J., Tian, H., McGuire, A.D., Kicklighter, D., Kittel, T., Rosenbloom, N., Running, S., Thornton, P., Ojima, D., Parton, W., Kelly, R., Sykes, M., Neilson, R., Brizzo, B., 2000. Contribution of increasing CO_2 and climate to carbon storage by ecosystems of the United States. Science 287, 2004–2006.

Shook, L.R., Tartal, J.A., 2000. Unreliable speculation or sound science under Daubert?—use of expert testimony to model contaminant release and transport. Am. Law Inst.-Am. Bar Assoc. 55, 525–545.

Sklash, M., Schroeder, M., Dragun, J., 1999. Groundwater models: Can you believe what they are saying? Nat. Resour. Environ. 13 (4), 542–546.

Smith, J.E., Heath, L.S., 2001. Identifying influences on model uncertainty: an application using a forest carbon budget model. Environ. Manage. 27 (2), 253–267.

Smith, J.E., Heath, L.S., 2002. A model of forest floor carbon mass for United States forest types. Research Paper NE-722. US Department of Agriculture, Forest Service, Northeastern Research Station, Newtown Square, PA, 37 pp.

Smith, J.E., Heath, L.S., Jenkins, J.C., 2003. Forest volume-to-biomass models and estimates of mass for live and standing dead trees of US forests. General Technical Report NE-298. US Department of Agriculture, Forest Service, Northeastern Research Station, Newtown Square, PA, 57 pp.

Stage, A., 2003. How forest models are connected to reality: evaluation criteria for their use in decision support? Can. J. For. Res. 33 (3), 410–421.

Soares, P., Tome, M., Skovsgaard, J.P., Vanclay, J.K., 1995. Evaluating a growth model for forest management using continuous forest inventory data. For. Ecol. Manage. 71, 251–265.

United Nations Framework Convention on Climate Change (UNFCCC), 1997. Kyoto Protocol to the United Nations Framework Convention on Climate Change. UNFCCC Document FCCC/CP/1997/7/Add. 1.

US Department of Agriculture, 2003a. Supplementary Guidelines for the Quality of Regulatory Information Disseminated by USDA Agencies and Offices. Office of the Chief Information Officer, Washington, DC. http://www.ocio.usda.gov/irm/qi_guide/regualtory.htm.

US Department of Agriculture, 2003b. Public Disclosure of Quality of Information Requests. Forest Service, Washington, DC. http://www.fs.fed.us/qoi/disclosure.shtml (accessed 12 December 2003).

US Environmental Protection Agency, 1994a. Guidance for Conducting External Peer Review of Environmental Regulatory Models (EPA 100-B-94-001). Office of the Administrator, Washington, DC. http://www.epa.gov/osp/spc/modelpr.htm.

US Environmental Protection Agency, 1994b. Validation strategy for the Integrated Exposure Uptake Biokinetic Model for lead in children. EPA 540/R-94-039.

US Environmental Protection Agency, 1999. White paper on the nature and scope of issues on adoption of model use acceptability guidelines. Science Policy Council, Model Acceptance Criteria and Peer Review White Paper Working Group. http://cfpub.epa.gov/crem/library/whitepaper_1999.pdf.

US Environmental Protection Agency, 2002. Guidance for Quality Assurance Project Plans for Modeling. EPA QA/G-5M. US Environmental Protection Agency, Washington, DC. http://cfpub.epa.gov/crem/library/QA_Guidance.pdf.

Vanclay, J.K., Skovsgaard, J.P., 1997. Evaluating forest growth models. Ecol. Model. 98, 1–12.

Versar Inc., 1988. Current and suggested practices in the validation of exposure assessment models. Report. Office of Health and Environmental Assessment, US EPA, Washington, DC.

Westerbrook, J.A., 1999. Air dispersion models: tools to assess impacts from pollution sources. Nat. Resour. Environ. 13 (4), 546–551.

Woodbury, P.B., Smith, J.E., Weinstein, D.A., Laurence, J.A., 1998. Assessing potential climate change effects on loblolly pine growth: a probabilistic regional modeling approach. For. Ecol. Manage. 107, 99–116.