iEMSs 2008:International Congress on Environmental Modelling and Software
Integrating Sciences and Information Technology for Environmental Assessment and Decision Making
4th Biennial Meeting of iEMSs, http://www.iemss.org/iemss2008/index.php?n=Main.Proceedings
M. Sànchez-Marrè, J. Béjar, J. Comas, A. Rizzoli and G. Guariso (Eds.)
International Environmental Modelling and Software Society (iEMSs), 2008

A Methodology for Building Credible Models for Policy Evaluation

C. A. Aumann [a]

[a]*Land-Use Management Systems, Alberta Research Council, 250 Karl Clark Road, Edmonton, Alberta, Canada (craig.aumann@arc.ab.ca)*

Abstract:

The need for tools capable of evaluating the potential impacts of alternative policies has been expressed by many. This paper focuses on a methodology that describes how credibility can be constructed for models used to evaluate alternative policies. Relative to modeling conducted in scientific contexts, however, modeling for policy evaluation has notable differences that would lead some to say that achieving credibility in such results is not possible. We first introduce a model assessment framework that enables us to describe how models for policy evaluation can still be more or less credible despite the differences from scientific modeling contexts. The argument presented depends primarily on i) the scalar hierarchical structure used to represent the complex policy system, ii) the ability of experimental frames to include a variety of constraints and/or weak data patterns across the scalar levels used to represent the system, and iii) the way in which the framework facilitates critique by stakeholders.

Keywords: Assessment; Critique; Complexity; Validation; Verification; Hierarchy theory; Simulation

1 INTRODUCTION

The need for tools capable of evaluating the potential impacts of alternative policies has been expressed around the world [e.g., Owens et al., 2004; Gov. Canada, 2005]. Despite this need, research over the last 30 years has demonstrated that the impact of policy assessment on policy decisions does not occur via a linear process - meaning that the outputs from policy assessments are generally not carefully considered or used directly by decision makers [Owens, 2005]. Instead, the impact of these tools occurs in more subtle and nuanced ways such as by facilitating group learning among stakeholders and providing ammunition that can be used to persuade opponents [Owens, 2005]. In considering how credibility can be established in the context of policy evaluation, the main conclusion is that while possessing credible models does not guarantee that a policy change will occur, possessing models that are credible to stakeholders and domain experts is indeed necessary for policy change to occur in contentious situations.

This paper thus focuses on the question of how to build credibility in models used to evaluate alternative policies. Others have described the issues associated with building such models [e.g., Couclelis, 2005; Jakeman et al., 2006]. After introducing a number of concepts and definitions to provide the necessary context associated with simulation modeling generally, a general assessment framework is described that builds on previous work [e.g, Zeigler et al., 2000; Aumann, 2007]. Some of the unique challenges associated with models used to evaluate policies are then discussed along with how the assessment framework presented can be applied to build credibility in policy evaluation contexts. Given that credibility must be established with stakeholders, the way in which results from this assessment framework could be delivered using web-based tools are also described.

To clarify the concepts and definitions introduced, a single running example is used throughout. This example focuses on how a model developed to evaluate a policy of tradable disturbance credits (TDCs) can be assessed. Similar to SO_2 emission trading, under TDCs a cap is placed on the cumulative amount of landscape disturbance that can occur within a region and industrial players within the region can purchase and trade permits for the right to disturb land [e.g., Weber and Adamowicz, 2002]. This is one policy under evaluation for efficiently achieving ecological and economic objectives in the oil-sands area of NE Alberta, Canada. An agent-based model is currently being used for this evaluation, but only a small number of the model's components are presented here to illustrate the concepts in this paper. The components used in this paper include industry agents (which attempt to maximize their economic return by exploring for oil across the landscape and use the exploration information to prioritize the drilling of wells and building of pipelines, etc), a government agent (which sets disturbance cap, monitors cumulative disturbance across the landbase, and auctions disturbance permits), and the natural environment (which includes things like natural disturbance and vegetative succession).

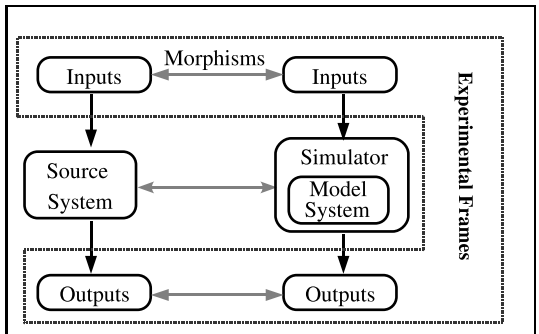

Figure 1: A general simulation framework consists of both source and model systems; the morphisms between the two systems in terms of input, outputs and model structure; and the experimental frames.

2 Assessment of Simulation Models

2.1 Simulation Framework

At a high level, a *model* is just a state transition function or mechanisms that instructs the *simulator* (e.g., a computer or algorithm capable of executing the model instructions) on how to generate outputs from inputs (Figure 1). The *source system* is the real or virtual system that is being modeled and the goal of modeling is to achieve a representation of the source system so that under "similar" inputs, both systems produce "similar" outputs. This idea is captured by the *experimental frames* which include a specification of the conditions under which both the source and model systems are to be observed or experimented with, along with mechanisms for comparing the two systems. The aim of model assessment is to build credibility in the model and is accomplished by ensuring that both systems "agree" over a sufficiently wide range of conditions encompassing the objectives motivating the modeling project.

One approach for representing complex systems is to decompose them according to a *scalar hierarchy* so that the objects at a given level contain, volumetrically and structurally, the objects of lower hierarchical levels [e.g., Kline, 1995; Giampietro, 2004]. As illustrated in Figure 2, one can think of the levels below a focal level as providing the "mechanism" while higher levels provide a "purpose" or context for the lower levels. Emergent properties are taken to define each level [Aumann, 2007].

At each level in this scalar hierarchy, model components can be specified with varying degrees of detail according to a *specification hierarchy*. An *I/O Behavior* specification is like a black-box in which inputs are mapped directly onto outputs. For example, in Figure 2 the way in which the Exploration component of the Oil Company is represented might follow an I/O behavior specification. In a *I/O System* specification, the model maintains an internal state that can be changed by model inputs. Thus, whether the Oil Company engages in exploration at a given time depends on it possessing sufficient TDCs. A *Coupled Component* specification is a model composed of other I/O Behavior, I/O System, and possibly other coupled component models and is illustrated by the entire Oil Company sub-model.

Scalar and specification hierarchies enable definition of what is meant by "similarity" between two systems [Zeigler et al., 2000]. Two systems are said to be *morphic* at a given specification level if it is possible to establish a direct correspondence between the defining elements of each system at the same specification level within some experimental frame. For an I/O Behavior specification, the morphism is simply the comparison of the inputs and outputs of both systems. For an I/O System specification, the introduction of a state space requires the introduction of a mapping between the state spaces of both systems. This mapping is said to be *homomorphic* if there is a defined correspondence (but not necessarily an identity) between the states in both systems so that both systems progress through similar pathways to achieve similar model outputs. Finally, since component model specifications can involve all three model specification types, it must be ascertained not only that the output of the overall model is correct (similar to I/O Behavior), but also that the outputs are produced for the right reasons (i.e., that the homomorphisms between the components of the two systems hold).

2.2 Model Assessment Framework

In the context of the simulation framework presented above, the goal of model assessment is to establish the strengths of the morphisms between the source and model systems. Building such credibility is accomplished via a processes of model verification, model validation, and critique. It should be noted that the meanings of these terms is not entirely consistent across fields with some eschewing the use of the term "validation" [e.g., Anderson and Bates, 2001], others noting the problems implied by the term while acknowledging it's widespread use [e.g., Oreskes and Belitz, 2001], and other distinguishing numerous different kinds of validation including operational, conceptual, data, and even processes [e.g., Rykiel Jr., 1996].

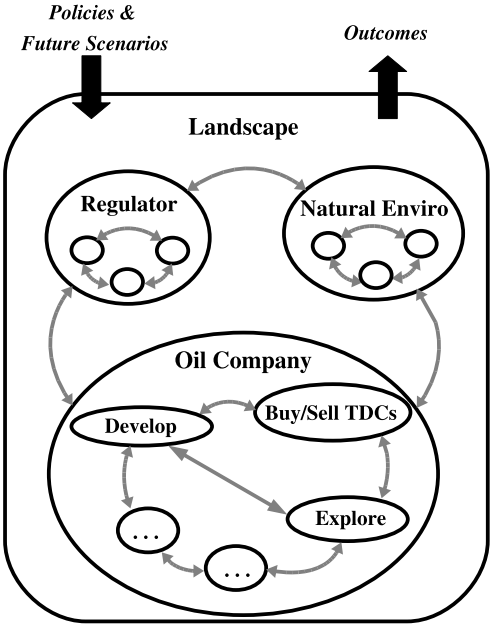

Figure 2: Simplistically, the entities relevant to evaluating the TDC policy include the oil companies, a regulator, and the natural environment. These entities all operate over a defined landscape area, and in turn are comprised of other lower-level components. In the case of an oil company, such components include units in charge of exploring for oil, developing oil (i.e., building roads and pipelines to get the oil to market), and also buying and selling TDCs with the Regulator or other companies.

In this paper, *model verification* means verifying that the actual model implementation is consistent with the model design specifications. Relative to the other assessment processes, verification is relatively easy to accomplish and thus this paper focuses on the processes of model validation and critique and builds on previous work [e.g, Balci, 1997; Rykiel Jr., 1996; Zeigler et al., 2000; Aumann, 2007].

Model validation is about substantiating that the behavior of the model "mimics" the behavior of the system with sufficient accuracy so that it is impossible to distinguish the behaviors of both systems in the experimental frames. *Experimental frames* are an operational formulation of the objectives motivating a modeling project and practically function as a type of measurement or observer system consisting of a *generator* that generates the input to the systems, an *acceptor* that monitors the "experiment" to ensure the desired experimental conditions are met by both systems, and a *transducer* that observes, analyzes and stores the output.

Saying it is impossible to distinguish the behaviors of two systems requires the concept of *replica-*

tive validity, namely that for all experiments possible within the experimental frame, the behavior of the model and the source system agree within the specified tolerance at the I/O Behavior level. For models specified at the I/O System level, the introduction of a state-space necessitates a stronger notion of *structural validity* meaning that the model mimics in a step-by-step, component-by-component fashion the way in which the source system performs its state transitions. Structural validity ensures that the model is generating the correct I/O behavior for the right reasons, and not because incorrect behaviors in one model component are compensated by behaviors in other model components. Methodologically, structural validity can be achieved by applying experimental frames to each of the model components within a given scalar level and across hierarchical levels to assess the larger scale consequence of these behaviors (see Figure 3 and discussion below).

One final, important step for achieving model credibility is a *critique* of the processes of model design, verification, and validation. Such critique is essential because model validity is only established relative to the study objectives as implemented in the experimental frames. If these objectives are incorrectly specified and/or the model is incorrectly defined, the model can still be valid with respect to this incorrect specification even though the simulation results will not be credible when viewed from a broader perspective. However, for such a critique to occur the model verification and model validation steps must be accessible, transparent, and understandable to non-modelers. How this can be achieved in the context of the current simulation and assessment framework will be discussed below.

We will say that a model is credible in a particular problem context if it has been verified, validated, and critiqued. No model can be absolutely credible, but rather models should be thought of in terms of degrees of credibility. Ultimately, the greater the need for high levels of credibility, the higher will be model development and assessment costs. Thus, the level of credibility required for the project needs to be bounded before the model is constructed and model assessment performed.

3 BUILDING CREDIBILITY IN MODELS FOR POLICY EVALUATION

While the above simulation and assessment frameworks appear quite natural for most systems studied by the physical sciences, a number of substantial differences arise in the context of policy evaluation that impact the way in which assessment can be carried out and ultimately the kind and level of credibility that is achievable. While models in scientific contexts are about constructing scientific explanations based on facts and well established theories, in a policy context the purpose is to decide on a course of action based on our values about the kind of unknown (and unknowable) future we desire based on knowledge that is always incomplete [Couclelis, 2005]. In addition, while scientific modeling will likely only ever be critiqued by a relatively small scientific community, a much larger community of diverse stakeholders will seek to comment on any policy assessment - especially if the financial stakes are high. These differences are explored more fully in this section.

3.1 Differences in the Simulation Framework

A major difference between typical scientific modeling and policy evaluation is that no source system exists for the later - since the point of such evaluation is to do it **before** a policy is ever enacted. Evaluation of a policy's performance must thus be judged relative to the future and not relative to observed behaviors - as is typical in scientific contexts. This unknown future is typically expressed using a number of alternative, coherent, plausible and relevant future scenarios capturing the exogenous conditions under which the policy might operate [e.g., Couclelis, 2005]. Thus, relative to the framework shown in Figure 1, the inputs to the model system also contain a set of future scenarios. The non-existence of the source system coupled with the alternative futures used means that limited data will exist for model validation, particularly at higher scalar levels that are critical to evaluating policy performance (e.g., experimental frame C in Figure 3). These differences help explain why the kind of credibility possible in scientific contexts cannot exist for policy evaluation.

3.2 Model Assessment - Validation

Despite these differences and their attendant challenges, we will argue that it is still possible for results of a policy evaluation to be more or less credible. The main tenets of the argument are based on the utility of constraints and "weak data patterns" when applied across the scalar levels of the model to validate model components, and the ability to apply experimental frames across multiple models.

In the context of the running example considered here, the impacts of a policy change would ideally be assessed regionally (in terms of economic and ecological indicators) and at lower scalar levels to capture the impacts on individual firms. The lack of empirical data may lead some to conclude that achieving any kind of credibility is impossible. However, such a position overstates the necessity of data in achieving model credibility. In general, the ability of any model to mimic data at the I/O Behavior level is only weak evidence that the important underlying processes have been correctly captured since models that inaccurately capture underlying processes can still predict quite well [Oreskes and Belitz, 2001]. For example, statistical models can predict quite well provided they are applied in a similar context to which they were built.

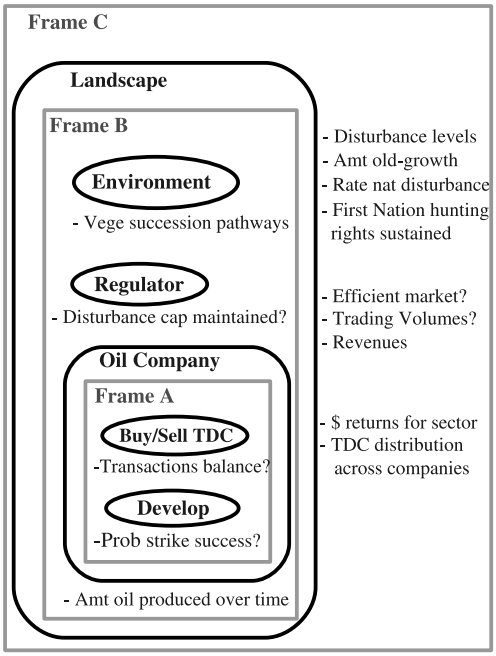

Figure 3: Experimental frames (in gray) encapsulate the lower-level components in Figure 2 where each frame contains criteria on the behavior of the sub-model that need to be respected. For example, the sub-model dealing with buying and selling TDCs needs to ensure that the transactions balance across the entire company, that the prices it is paying for the TDCs are not causing the company's financial ruin, etc. At higher hierarchical levels, the amount of oil produced by a company over time should not fluctuate widely, the market should be efficient, and overall trading volumes for TDCs should be large given a large number of players in the market.

In the context of a policy that alters any existing context for which data may exist, validity can be established using structural integrity criteria at the I/O System level. Structural integrity criteria ensure that the answers achieved at a given scalar level are also achieved for the "right reasons" - meaning that all of the experimental frames applied to lower levels are also satisfied. In the context of the scalar hierarchical model decomposition used, it must be decided whether the behaviors of sub-models are the same under the new policy context. For example, under TDCs the vegetation succession pathways will remain the same, as should the probability of a strike success as exploration proceeds across a basin (see Figure 3). In this case, existing data can be used.

For sub-model behaviors that change under alternative policies, the experimental frames can include a number of alternative options: i) constraints can be applied to these components (e.g., Do a company's number of TDCs and the disturbance it creates balance? Is the disturbance cap being maintained across the landscape?); ii) qualitative behavior assessment criteria (e.g., Are companies developing in areas of actual high petroleum potential? Are companies generally selling TDCs to each other when it makes sense to do so? Is the market efficient?); and iii) "vague" or "weak" data patterns (e.g., Do the natural disturbance patterns created by the model agree with what is observed empirically?). Thus, the lack of "hard data" for the system being modeled is no excuse for failing to perform model validation in policy contexts.

Applying such types of "data" across the scalar hierarchical levels using the experimental frames can provide as much information for model validation as a single strong empirical data pattern [Grimm et al., 2005, & Online Material] and also [Aumann et al., 2006, Appendix B]. The reason for this is that **all** experiments specified for the scalar hierarchical levels below the current level must be satisfied **simultaneously**. Provided the criteria used in the experimental frames are not trivial, satisfying all frames simultaneously rapidly becomes challenging. Further, the way in which experimental frames trade-off with each other can be used to guide refinement of model structure and also refine the experimental frames [Reynolds and Ford, 1999; Wiegand et al., 2003].

Another way in which model validity can be established is by comparing the behaviors of alternative models constructed using different modeling assumptions. In the context of such multiple models, the experimental frames are general enough to allow these models to be validated against each other. However, the challenge here is ensuring that the experimental frames are *applicable* - meaning that the conditions required by an experimental frame can be satisfied by all the models.

In summary, the non-existence of the source system and the lack of data does not mean model validation can be ignored nor that validation is impossible. Instead, the above validation process allows us to say why the model's behaviors are being produced and to demonstrate that these behaviors are being produced for the right reasons. The reason we are justified in believing the outcomes produced by the model results from the behavior of the lower level model components being deemed to be valid based on the experimental frames applied at these level(s) coupled with the experimental frames applied across higher hierarchical levels. As a result, we can have confidence that these lower levels are not spuriously influencing the behavior of higher levels, and that the behaviors of model components are all within reasonable bounds. Because the outputs are produced for reasons that we think are justifiable (if we didn't think this, then we would add additional experimental frames or include models built using alternative assumption), we are compelled to view the results as valid within the assumptions of the assessment.

This approach does, however, also have some notable limitations. First, even if alternative models are constructed and these models all agree across the applicable experimental frames, there is a very real possibility that all of such models are simply wrong due to a common lack of knowledge about the non-existent system being modelled. Thus, agreement across diverse models may simply be a result of common ignorance. While disagreement across models might improve understanding at the stage of critique, this is unlikely to occur under common ignorance and is particularly challenging since no source system exists to act as the ultimate authority. Another limitation in the above method is that evaluations can only be done under a small and finite set of alternative futures. While utilizing a greater variety of alternative futures will help to ensure that a "robust policy" is identified [e.g., Lempert et al., 2006], constructing a large number of alternative futures that are coherent, plausible and realistic presents its own challenges.

3.3 Model Assessment - Critique

A larger critique of model design and validation is essential because model validity is a necessary but insufficient condition for establishing the credibility of modeling results [e.g., Balci, 1997]. As is clear from the discussion above, model validity is only established relative to the experimental frames used. If these experimental frames are inadequate for the study objectives, the model can still be deemed to be valid even though the simulation results will not be credible. Credibility can be enhanced via a process of external review by stakeholders. However, the challenge is delivering the large quantities of information produced during validation in a manner that is accessible, transparent, and understandable to such reviewers who are likely not modelers. This section illustrates conceptually how such delivery could be accomplished under the simulation framework presented.

Since the system being modeled is conceptualized in terms of a scalar hierarchy, these entities are represented as separate sub-models within the overall model. Each of these sub-models is also "wrapped" (e.g., Figure 3) in one or more experimental frames that may encompass one or more

sub-models. Each experimental frame has associated with it criteria governing the inputs that are allowable for the sub-model (generator), criteria to ensure that the assumptions underlying the sub-model are maintained (acceptor), and the frame also monitors the output produced by the sub-model to ensure its acceptability (transducer). Problems could originate in any of these three areas, and the aim is to present such failures to the user in an easy to understand way and enable them to browse the criteria used in the experimental frames. One possibility is illustrated in Figure 4.

Current programming tools enable the hierarchical structure of models to be displayed. Thus, the challenge is to expand these tools to display the experimental frames and also add in an overarching monitoring system that monitors and reports on the status of all the frames as the model runs. Conceptually, users could browse such information as illustrated in Figure 4. The challenge is implementing such a tool in a general manner so that it can deal with models from diverse problem contexts.

4 CONCLUSION AND RECOMMENDATIONS

With the ever increasing power of computers, the amount of modeling done and the complexity of the models built in both scientific and policy fields will continue to increase. However, to date the field of model assessment lags far behind our current computational abilities. Indeed, much work is still needed to arrive at the concepts and a framework for model assessment that is general across disparate disciplines. This paper attempts to advance the field of model assessment, at least as it relates to policy assessment, by introducing a number of concepts that have proven useful in scientific modeling contexts and describing how these same concepts apply to policy evaluation.

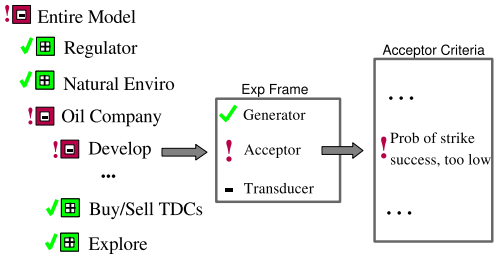

Figure 4: Using the hierarchical structure of the model and the fact that experimental frames cover different sub-models, it becomes possible for a user to quickly identify which sub-models are failing and why, and communciate these results to stakeholders.

The strengths of this framework are that it can be applied in both scientific and policy contexts, suggesting that it is at least somewhat general. Conceptually, the framework lends itself to the communication of model validation results along with the experimental frames in a manner that is accessible, transparent, and likely understandable to stakeholders (Figure 4). Drawbacks to the framework include that implementing it will require substantially more programming effort and running it along with the model will require considerable computational resources.

This framework would seem to be a natural starting place for enabling autonomic model assessment tools - meaning tools that are capable of assisting in the complex process of model validation and the identification of inadequate model components or experimental frames. While autonomic computing aims to create computer systems that are capable of self management - the processes of model verification, model critique, and the specification of the experimental frames all require levels of human involvement that go far beyond anything a computer can do for us. In a policy context where one of the aims of modeling is to foster dialog among stakeholders, we do not see the inability to fully automate this model assessment framework as a drawback. Instead, a more urgent research need is figuring out how to best incorporate model assessment tools into stakeholder consultation to facilitate public consensus on which policy is prefered.

REFERENCES

Anderson, M. G. and P. D. Bates. Hydrological science: model credibility and scientific understanding. In Anderson, M. G. and lastname Bates, P. D., editors, *Model Validation: perspectives*

in hydrological science, chapter 1, pages 1–10. John Wiley & Sons, New York, 2001.

Aumann, C. A. A methodology for developing simulation models of complex systems. *Ecological Modelling*, 202:385–396, 2007.

Aumann, C. A., L. A. Eby, and W. F. Fagan. How transient patches affect population dynamics: the case of hypoxia and blue crabs. *Ecological Monographs*, 76(3):415–438, 2006.

Balci, O. Principles of simulation model validation, verification, and testing. *Transactions of the Society for Computer Simulation International*, 14(1):3–12, 1997.

Couclelis, H. "where has the future gone?" rethinking the role of integrated land-use models in spatial planning. *Environment and Planning A*, 37:1353–1371, 2005.

Giampietro, M. *Multi-scale Integrated Analysis of Agroecosystems*. CRC Press, New York, 2004.

Gov. Canada. Integrated landscape management models for sustainable development policy making. `http://www.policyresearch.gc.ca/doclib/ILMM_Workshop_Report_e.pdf`, 2005.

Grimm, V., E. Revilla, U. Berger, F. Jeltsch, W. M. Mooij, S. F. Railsback, H. Thulke, J. Weiner, T. Wiegand, and D. L. DeAngelis. Pattern-oriented modeling of agent-based complex systems: lessons from ecology. *Science*, 310(11):987–991, November 2005.

Jakeman, A. J., R. A. Letcher, and J. Norton. Ten iterative steps in development and evaluation of environmental models. *Environmental Modelling & Software*, 21:602–614, 2006.

Kline, S. J. *Conceptual Foundations for Multidisciplinary Thinking*. Stanford University Press, 1995.

Lempert, R., D. g. Groves, S. W. Popper, and S. C. Bankes. A general, analytic method for generating robust strategies and narrative scenarios. *Management Science*, 52(4):514–528, 2006.

Oreskes, N. and K. Belitz. Philosophical issues in model assessment. In Anderson, M. G. and lastname Bates, P. D., editors, *Model Validation: perspectives in hydrological science*, chapter 3, pages 23–41. John Wiley & Sons, New York, 2001.

Owens, S. Making a difference? some perspectives on environmental research and policy. *Transactions of the Institute for British Geographers*, 30:287–292, 2005.

Owens, S., T. Rayner, and O. Bina. New agendas for appraisal: reflections on theory, practice, and research. *Environment and Planning A*, 36:1943–1959, 2004.

Reynolds, J. H. and E. D. Ford. Multi-criteria assessment of ecological process models. *Ecology*, 80(2):538–553, 1999.

Rykiel Jr., E. J. Testing ecological models: the meaning of validation. *Ecological Modelling*, 90: 229–244, 1996.

Weber, M. and V. Adamowicz. Tradable land-use rights for cumulative environmental effects management. *Canadian Public Policy*, 28(4):581–595, 2002.

Wiegand, T., F. Jeltsch, I. Hanski, and V. Grimm. Using pattern-oriented modeling for revealing hidden information: a key for reconciling ecological theory and application. *Oikos*, 100:209–222, 2003.

Zeigler, B. P., H. Praehoffer, and T. G. Kim. *Theory of Modeling and Simulation*. Academic Press, New York, 2 edition, 2000.