Research Repository UCD

Title	Technical assessment and evaluation of environmental models and software : letter to the Editor
Authors(s)	Alexandrov, G. A., Ames, D., Bellocchi, G., Bruen, Michael, Crout, N., Erechtkoukova, M., Hildebrandt, A., Hoffman, F., Jackisch, C., Khaiter, P., Mannina, G., Matsunaga, T., Purucker, S. T., Rivington, M., Samaniego, L.
Publication date	2011-03
Publication information	Alexandrov, G. A., D. Ames, G. Bellocchi, Michael Bruen, N. Crout, M. Erechtkoukova, A. Hildebrandt, et al. "Technical Assessment and Evaluation of Environmental Models and Software : Letter to the Editor" 26, no. 3 (March, 2011).
Publisher	Elsevier
Item record/more information	http://hdl.handle.net/10197/3064
Publisher's statement	This is the author's version of a work that was accepted for publication in Environmental Modelling & Software. Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published in Environmental Modelling & Software Volume 26, Issue 3, March 2011, Pages 328-336 DOI: 10.1016/j.envsoft.2010.08.004
Publisher's version (DOI)	10.1016/j.envsoft.2010.08.004

Downloaded 2023-04-07T10:52:29Z

The UCD community has made this article openly available. Please share how this access benefits you. Your story matters! (@ucd_oa)

© Some rights reserved. For more information

1 **[Published in Environmental Modelling and Software]**

2 **Technical assessment and evaluation of environmental models**
3 **and software: Letter to the Editor:**

4
5 Alexandrov, G.A.^{1*}, Ames, D., Bellocchi, G.⁴, Bruen, M.⁵, Crout, N.⁶, Erechtkhoukova,
6 M.⁷, Hildebrandt, A.⁸, Hoffman, F.⁹, Jackisch, C.¹⁰, Khaiter, P.⁷, Mannina, G.¹²,
7 Matsunaga, T.¹, Purucker, S.T.¹⁴, Rivington, M.¹⁵, Samaniego L.⁸, Sherman, B.

8
9 ¹Center for Global Environmental Research, National Institute for Environmental
10 Studies, Tsukuba, Ibaraki, Japan

11
12 ⁴ Grassland Ecosystem Research Unit, French National Institute for Agricultural
13 Research, Clermont-Ferrand, France

14
15 ⁵UCD Centre for Water Resources Research, University College Dublin, Ireland

16
17 ⁶School of Biosciences, University of Nottingham, UK

18
19 ⁷ York University, Toronto, Canada

20
21 ⁸ UFZ-Helmholtz-Centre for Environmental Research, Leipzig, Germany.

23 ⁹ Oak Ridge National Laboratory, Computational Earth Sciences Group, PO Box 2008,
24 Oak Ridge, TN 37831, USA

25

26 ¹⁰ Technical University Munich, Munich, Germany.

27

28 ¹² Department of Hydraulic Engineering and Environmental Applications, University of
29 Palermo, Italy.

30

31 ¹⁴ US Environmental Protection Agency, Athens, Georgia, USA.

32

33 ¹⁵ Macaulay Land Use Research Institute, Craigiebuckler, Aberdeen. AB15 8QH. United
34 Kingdom.

35

36

37

38 *Corresponding author g.alexandrov@nies.go.jp

39

40

Rationale

This letter details the collective views of a number of independent researchers on the technical assessment and evaluation of environmental models and software. The purpose is to stimulate debate and initiate action that leads to an improved quality of model development and evaluation, so increasing the capacity for models to have positive outcomes from their use. As such, we emphasise the relationship between the model evaluation process and credibility with stakeholders (including funding agencies) with a view to ensure continued support for modelling efforts.

Many journals, including EM&S, publish the results of environmental modelling studies and must judge the work and the submitted papers based solely on the material that the authors have chosen to present and on how they present it. There is considerable variation in how this is done with the consequent risk of considerable variation in the quality and usefulness of the resulting publication. Part of the problem is that the review process is reactive, responding to the submitted manuscript. In this letter, we attempt to be proactive and give guidelines for researchers, authors and reviewers as to what constitutes best practice in presenting environmental modelling results. This is a unique contribution to the organisation and practice of model-based research and the communication of its results that will benefit the entire environmental modelling community. For a start, our view is that the community of environmental modellers should have a common vision of minimum standards that an environmental model must meet. A common vision of what a good model should be is expressed in various guidelines on Good Modelling Practice. The guidelines prompt modellers to codify their practice and to be more rigorous in their model testing. Our statement within this letter deals with another aspect of the issue - it

prompts professional journals to codify the peer-review process. Introducing a more formalized approach to peer-review may discourage reviewers from accepting invitations to review given the additional time and labour requirements. The burden of proving model credibility is thus shifted to the authors. Here we discuss how to reduce this burden by selecting realistic evaluation criteria and conclude by advocating the use of standardized evaluation tools as this is a key issue that needs to be tackled.

Background

The use of models for any practical purpose entails the risk of misuse. If a model's limitations are not completely understood, the model outputs may be easily misinterpreted (Jakeman et al., 2009). To reduce this risk, every model should be assessed and evaluated by domain experts - that is, by modellers experienced in model development and application.

Such assessment and evaluation is normally undertaken when an article describing a model (or software) passes through a peer-review system of a professional journal (Fig 1) such as EM&S, for example. Most experienced modellers are involved in the peer-review process and periodically evaluate models made by their colleagues. Therefore, the community of environmental modellers needs to have a common vision of minimum standards that an environmental model must meet.

A common vision of what a good model is has been expressed in guidelines on Good Modelling Practice (STOWA/RIZA, 1999; Murray-Darling Basin Commission, 2000; Jakeman et al., 2006; Gaber et al., 2008; Robson et al., 2008; Welsch et al., 2008). The guidelines prompt modellers to codify their practice and to be more rigorous in their model testing. The purpose of this letter is to deal with another aspect of the issue: that of prompting professional journals to codify the peer-review process. The objective is to highlight the obstacles to model evaluation and to provide possible solutions. This is not however a review on the state of model evaluation, the latter being given in a recent

paper by Bellocchi et al. (2010). Rather we seek to promote improvement within the quality of model evaluation as part of the peer-review process. In doing so, we have also highlighted issues that potential reviewers need to be aware of.

Peer review is normally considered as an essential component of research dissemination and remains the principal mechanism by which the quality of research is judged (Council of Science Editors, 2006; Müller, 2009). At the same time, there is common understanding that peer-review cannot be expected to detect fraud and ensure perfection (Hames, 2007): *“even the most-respected journals have been caught out and, despite extensive peer review, have ended up publishing fraudulent or seriously flawed material”* (Wager, 2006). Then, what is the main purpose of peer-review? There is no general agreement on this issue now. One may suggest that the peer-review system initially introduced for filtering out unreasonable claims to new research results still serves this purpose (Walker, 1998; Alexandrov, 2006).

In the case of complex models, it is likely the process will result in reviews not evaluating the components of the model (especially if referenced to other sources), on the defence that journal readers and end users who are specialists on those components will make their own judgments. In reality, reviewers will not have sufficient time or resources to conduct detailed evaluation of individual components or the whole model let alone the software coding. Hence, the emphasis must be on model developers to provide accurate evidence, covering sufficient complexity interactions, to demonstrate adequate testing to achieve a stated level of model reliability and utility. We feel this requires clearer

statements (supported by evidence) from the model developers on known limitations and areas of uncertainty. Such an open approach should, if communicated correctly, i.e. positively through addressing the consequences of any uncertainty, actually increase credibility with end users rather than diminish it. This is important as perceptions of uncertainty and how it is handled amongst researchers, policy makers and politicians have changed recently, especially since the rise of climate change modelling and planning for adaptation. Previously end users (particularly policy makers and politicians) were reluctant to deal with the realities of the uncertainty associated with modelling. Reduction of uncertainties in the models may be pursued by, for instance, model-data integration techniques (e.g. Wang et al., 2009). However, it is clear that the skill in handling uncertainty not only lies within statistical and other forms of model testing, but also in how it is communicated to end users. Hence, our view that the establishment of a standardized set of criteria and methods for model evaluation is needed to set a minimum standard for 'proof of testing' that would serve to support uncertainty communication. The absence of such standardized criteria and methods risk modelling becoming unacceptable as a form of research for predictive purposes.

The major obstacles to a more formalized approach to model evaluation

Currently there is no requirement to detail a full set of model specifications. The first task in standardization therefore would be to have a scheme whereby published models could

141 be fully specified (Fig 2). Depending on the rationale for the research exercise, modelling
142 for theoretical scientific purposes and modelling for decision-making may follow
143 separate paths (Haag and Kaupenjohann, 2001) and hence require different specifications
144 for assessment and evaluation. In general, this would require a definition of the modelling
145 objective, its formulation, implementation and parameterization, further supplemented by
146 information on how they have been evaluated and the conclusions of that process.

147
148 A standardized set of criteria with which models should be assessed and evaluated will at
149 least ensure the minimum of review effort is made. The risk, however, is that a more
150 formalized approach to peer review requiring the achievement of a minimum standard
151 may discourage reviewers from accepting invitations to review given the additional time
152 and labour requirements. Further to this, a limitation of past model development and
153 application has been that funding organizations have been reluctant to accept the
154 additional costs of performing appropriate model assessment and evaluation within
155 proposals from researchers. Hence, model development has often been on tight budgets
156 causing assessment and evaluation to take a lower priority. Specification of minimum
157 standards for assessment and evaluation and how it is reported should hence also form the
158 basis for the minimum level of validation effort written into funding proposals. Similarly,
159 organizations awarding grants need to include in their calls for proposals more explicit
160 details on the requirements for evaluation and testing and be prepared to provide the
161 required funds.

163 This implies that funding organizations have to take on board a greater level of
164 responsibility in supporting modelling work that includes increased assessment and
165 evaluation efforts. Funders could effectively impose minimum standards for model
166 development, evaluation and specification, and progress could be made in this area if
167 ‘lead’ funders introduced such requirements. For example, in the United Kingdom
168 context if the national research councils were to introduce minimum standards other
169 funders would follow in time.

170
171 A further obstacle to overcome lies within the community of environmental modellers
172 itself, which has to take on board a greater level of responsibility in developing standards.
173 The procedures to perform the evaluation task are not widely accepted (Cheng et al.,
174 1991) and appear in several forms, depending on data availability, system characteristics
175 and researchers’ opinion (Hsu et al., 1999). Environmental models are made up of
176 mixtures of rate equations, comprise approaches with different levels of empiricism, aim
177 at simulating systems which show non-linear behaviour and often require numerical
178 rather than analytical solutions. Therefore, the computer program, including technical
179 issues and possible errors, is tested rather than the mathematical model representing the
180 system (Leffelaar et al., 2003). Hence, given the applied nature of models in representing
181 a system, their usefulness can be evaluated only in specific case studies. Gardner and
182 Urban (2003) suggested assessing model usefulness based on its appropriateness and
183 performance. Model appropriateness describes the extent to which the model meets the
184 objectives of the study. The appropriateness usually deals with the model structure,
185 although the necessity to identify model parameters brings observation data into the scene

(e.g. Confalonieri et al., 2009b). The availability or unavailability of observational data largely predetermines the structure of a model. Model performance is evaluated based on reported testing results in such terms as “goodness of fit” between simulated values of model variables and observation data and required computational time. Our observation is that quantification of uncertainty is less often reported.

Evaluation of model uncertainty is an important part of model assessment, yet application specific since it depends on model parameterization. On different sets of parameter values, the same mathematical equations may exhibit substantially different dynamic features. Thus, in dynamic models, changes in model parameters can trigger a switch from stable solution to an unstable one, causing a significant increase in model uncertainty (e.g. van Nes and Scheffer, 2003). Stability analysis of a solution must be a part of model investigation, but the analysis may become complicated for complex models, and has therefore not often been undertaken.

However, we recognise that a complete evaluation of model uncertainty is hardly possible. Usually, the analysis is confined to quantifiable sources, such as initial values of state variables and parameters. Indeed as pointed out by Harremoës (2003) not all uncertainty sources can be ‘quantified’, and that the fraction of uncertainty source terms being ‘ignored’ might be high in environmental investigations. The investigation of the model structural uncertainty is uncommon. Even if an estimate of uncertainty is obtained, its interpretation is not straightforward. The term ‘high uncertainty’ is ambiguous and was defined rather intuitively. Reichert and Borsuk (2005) considered the uncertainty as

‘high’ when the width of predicted distribution of model solutions is larger than the difference between expected outcomes of different simulated alternatives. Strictly speaking, an absolute value of the uncertainty is not important as long as simulations allow for a clear distinction between considered scenarios and for comparison of projected outcomes against some known objectives. In other words, the interpretation of model uncertainty is also application dependent. Codifying the testing process by the model authors will establish an uncertainty range of at least one application case.

It is often stated that a clear understanding of the model’s purpose is central to its evaluation. In other words, it should be ‘fit for purpose’. In fact the use of this phrase can be helpful, as it places the purpose ‘up front’ and emphasizes that generally perfection is not sought, merely the fitness for the given purpose. It is important to distinguish between cases where the purpose is prediction to underpin a decision-making process, and those cases where the model serves as a test bed for scientific hypotheses (even though the same underlying model may be used in both situations). In the decision support case, the accuracy of the predictions for the given purpose is important, as are other factors such as the input data requirements, the safe operating domain of the model and stakeholders’ acceptance of the model. In the scientific method case, the evaluation generally needs to be more sophisticated. It is not enough to confirm the hypotheses contained in the model based on the agreement between predictions and observations. A further test is required to rule out the possibility that alternative model formulations (i.e. different hypotheses) could also have described the observations available. This relates to the ‘equifinality’

thesis of Beven and Freer (Beven and Freer, 2001; Beven, 2006) and the issue of choices in model formulation (e.g. Cox et al., 2006; Crout et al., 2009).

The solutions: realistic criteria for model evaluation

Since the range of modelling situations is wide, we recognise that generally applicable standards can be formulated only in a generic form. They form a framework for model evaluation leaving the details of a particular implementation, such as quantification of criteria, up to the reviewers. Starting with the technical assessment of a model, we suggest reviewers first answer the questions below and evaluate the developers' claim on the usefulness of an environmental model:

- Do developers delineate the domain of model application?
- Do they highlight advanced model features against the prior art?
- Do they provide an example of model application illustrating model performance?

Then, they may proceed to assessment of the “proofs” of model usability, which are expected to show that:

- The domain of model application is delineated correctly;
- The model has certain advantages over a prior art;
- The example of model application shows credibility of the model as a tool for environmental assessment.

In the next sections, we expand our views on the above points.

251 ***How to delineate the domain of model application***

252 An environmental model is normally developed using a four-tier approach: conceptual
253 scheme, model formulation, computer code, and specific parameterization. Consequently,
254 delineating the domain of its applicability one should clearly make a distinction between
255 applicability of each tier. A conceptual scheme may be applied over a large range of
256 environmental states, whereas its specific parameterization may be intended for use under
257 very restrictive conditions. In addition, the model code may be suitable for use only
258 within a certain range of model parameters and inputs.

259

260 The four-tier description of model domain should answer the following questions:

261

- 262 • Which environmental states may fall within the conceptual scope of the model?
- 263 • Which environmental states may be assessed (or explored) using the current
264 version of a model in question or its computer code?
- 265 • Which environmental states may be assessed (or explored) using a specific
266 parameterization of the model?

267

268 The conceptual scheme of a model is derived from the model developer's perceptual
269 model of the real system at hand. The perceptual model is known to be an approximation
270 (to a greater or lesser extent). Moreover, it is common to have a range of scientific
271 opinions regarding the best representation of the perceptual model. Nevertheless, it is
272 always possible to make the distinction between environmental conditions that may fall
273 within the conceptual scope of the model and those that may not. For example, if the

conceptual scheme of a model does not address some environmental factors, the model may not assess the environmental impact of this factor.

In general, the correct description of the model domain must guarantee that the model will not produce results that go beyond empirically (or theoretically) established bounds. A related part of the technical assessment is to find the “regions” of the declared model domain, where the model produces obviously erroneous results, or confirm that no such “regions” were found. The latter helps to evaluate model reliability defined by Mankin et al. (1975).

How to show that a model has an advantage over the prior art

The purpose of developing a new model is to make visible progress in the state-of-the-art (Jørgensen et al., 2006). This can be done in different ways. The simplest of them is improving either the conceptual scheme or computer code of a prior model. In this case, the advantage over the prior art may be highlighted by providing some proof that:

- The model addresses environmental situations that do not fall within the scope of the prior model(s);

or that:

- The model code is more efficient than that of the prior model(s) (e.g., needs less initial information) in addressing some environmental situations;

or that:

- The specific parameterization of the model shows better performance than that of the prior model(s) in addressing specific environmental conditions.

297

298 In the well-developed fields of environmental modelling, the multi-model approach is
299 considered to be more reasonable than the best-model approach. Multi-model
300 combinations outperform best models. In other words, the progress in the state-of-the-art
301 is achieved through improving performance of a multi-model ensemble.

302

303 The examples of testing in such cases include multi-model analysis (MMA) for
304 developing multiple plausible models by considering alternative processes, using
305 alternative modelling codes, or by defining alternative boundary conditions (Pachepsky et
306 al., 2006). Quantitative MMA methods assign performance scores to each candidate
307 model (e.g. Burnham and Anderson, 2002; Ye et al., 2008). The scores are utilized to
308 rank and select the best models or to assign importance weights (e.g., for use in an
309 ensemble forecasting). Qualitative MMA methods can also rely on expert elicitation,
310 stakeholder involvement, and quality assurance/quality control procedures to assess
311 relative merits of alternative models (Funtowicz and Ravetz, 1990; van der Sluis, 2007).

312

313 Improving the mathematical formulation of a given conceptual scheme is also a way for
314 improving the state-of-the art. The selection of a suitable formulation relates to model
315 comparisons that cannot be fully ‘automated’ or formalized due to a confounding effect.
316 Confounding appears when two or more factors cause a combined measurable effect and
317 the contribution of each individual factor cannot be estimated separately. Thus, a
318 particular value of a model parameter depends not only on the corresponding state
319 variable and processes included in the model, but also on a given formula used to

320 describe each process. The majority of environmental models require a number of
321 parameters that must be identified for a given case study. In such a case, the comparison
322 of different models becomes dubious because it is hard to differentiate (in the overall
323 model uncertainty) the effect created by model structure from the effect generated by the
324 assigned values of model parameters.

325
326 Moreover, even a small change in a sub-model introduced to correct its functionality may
327 produce a different interpretation on simulated processes. The reason for these unwanted
328 changes lies in the lack of independence/wrong dependencies of parts of the code, which
329 is not completely avoidable. This aspect might go beyond a simple evaluation by once
330 again comparing against previously acceptable results (Huth and Holzworth, 2005) and
331 poses the need for formal model evaluation against observed data at each published stage
332 of model development (van Oijen, 2002). Each version of a model, throughout its
333 development life cycle, should be subjected to output testing, designed by identifying test
334 scenarios, test cases, and/or test data.

336 ***How to show model credibility***

337 The establishment of credibility is a prerequisite for model acceptance and use.
338 Credibility is in itself a complex issue extending beyond just model testing (e.g
339 authenticity of problem ownership, skills and motivation of the research team developing
340 a model, etc.). Model evaluation is, however, the key starting point for establishing
341 credibility. Hence, a strengthened peer review procedure will have an essential role in the
342 credibility building process. However, model evaluation must not be seen as a one-off

event or a “once-and-for-all” activity (Janssen and Heuberger, 1995), but as an on-going process to check for model compatibility to current evidence and variations (e.g. in spatial, climatic and hydrological conditions). Moreover, according to Sinclair and Seligman (2000), demonstration that a models’ output more or less fits a set of data is a necessary but not sufficient indication of validity. This is because model validity is rather the capability to analyze, clarify, and solve empirical and conceptual problems. Empirical problems in a domain are, in general, about the observable world in need of explanation because a model does not adequately solve it, rival models solve it in different ways, or it is solved/unsolved depending on the model. Conceptual problems arise when the concepts within a model appear to be logically inconsistent, vague and unclear, or circularly defined, and when the definition of some phenomenon in a model is hard to harmonize with an ordinary language or definition (e.g. Parker, 2001). This raises the issue of widening beyond numerical testing by also including stakeholders’ evaluation and expert interpretation through soft systems approaches (Bellocchi et al., 2002; Matthews et al., 2008). For example, non-scientific end users may be more persuaded of model validity by graphical representations than statistical tests or indices, especially where historical events or familiar phenomena are shown and are recognizable by them.

Thus, to evaluate a model as a credible one, a reviewer should confirm at least that:

- Its conceptual scheme is theoretically adequate to the declared domain of applicability;
- Its computer code is verifiable;
- The accuracy of its specific parameterization is consistent with intended usage.

366

367 **Adequacy and prediction**

368 The model adequacy cannot be assessed regardless of the domain of its applicability
369 (Rykiel, 1996). The context within which models are used affects the required
370 functionality and/or accuracy (French and Geldermann, 2005). This is particularly
371 apparent when comparing models developed to represent the same process at different
372 scales and for which different qualities of input, parameterization and validation data will
373 be available, for example soil water balances at plot, farm, catchment and region (e.g.
374 Keating et al., 2002; Viscel et al., 2007). This has led to the development of application
375 specific testing of models and the idea of model benchmarking, by comparing simulation
376 outputs with outputs of another simulation that is accepted as a “standard” (e.g. Vanclay,
377 1994). Such approaches typically use multi-criteria assessment (e.g. Reynolds and Ford,
378 1999) with performance criteria weighted by users depending on their relative
379 importance.

380

381 Such indications of adequacy are essential in relation to the use of models for future
382 predictive purposes. Papers on modelling often state that they aim to produce an
383 instrument for prediction (van Oijen, 2002). A fundamental issue is to quantify the degree
384 to which a model captures an underlying reality and predicts future cases (Marcus and
385 Elias, 1998; Li et al., 2003). Predictions pose special problems for testing, especially if
386 prediction focuses on events in the far future. Predictive models can be accepted if they
387 explain past events (*ex-post* validation). The probability of making reasonable projections
388 decreases with the length of time looked forward. A continuous exchange of validation

data among developers and test teams should either ensure a progressive validation of the models by time, or highlights the need for updated interpretations of the changed system.

In many cases, predictive models are mixed with exploratory models. The distinction between them can be drawn on the basis of data availability. Predictive models are normally used in connection with an observing system established for environmental monitoring. Exploratory models, in contrast, are normally used where observations are limited.. Therefore, testing methods need to be appropriate for each case, in order to demonstrate adequacy for each purpose.

Code verifiability

Computer code is a translation of mathematical clauses from the mathematical language to a computer language. The one-to-one correspondence is not always achieved. There is some consensus (after Glasow and Pace, 1999) that component-based development is indeed an effective and affordable way of creating model applications and conducting model evaluation.

In such a case, it is our view that particular emphasis should be placed on designing and coding object-oriented simulation models to properly transfer simulation control between entities, resources and system controllers and on techniques for obtaining a correspondence between simulation code and system behaviour. It is crucial to consider the issue of model component validity when considering model re-use as it needs to be a fundamental part of any re-use strategy.

412

413 The distribution of already validated model components (mathematical and coded
414 algorithms) can substantially decrease the model validation effort when re-used. A key
415 step in this direction is the coupling between model components and evaluation
416 techniques, the latter also being implemented into component-based software. Such
417 evaluation systems should stand at the core of a general framework where the modelling
418 system (i.e. a set of modelling components) and a data provider supply inputs to an
419 evaluation tool (e.g. Bellocchi et al., 2006). Such an evaluation tool is also meant as a
420 component-based system, both communicating with the modelling component and the
421 data provider via a suitable protocol and allowing the user to interact in some way (e.g.
422 via a graphical user interface) to choose and parameterize the evaluation tools.

423

424 The output from an evaluation system can be offered to a deliberative process (e.g.
425 stakeholder review) for interpretation of results. Adjustments in the modelling system or
426 critical reviewing of data used to evaluate the model can be a next stage, if the results are
427 assessed as unsatisfactory for the application purpose. A new evaluation-interpretation
428 cycle can be run any time new versions of the modelling system are developed and
429 plugged in to the evaluation component. Again, a well-designed, component-based
430 evaluation system can be easily extended towards including further evaluation
431 approaches to keep up with evolving methodologies, e.g. statistical, neural networks or
432 fuzzy-based (e.g. Bellocchi et al., 2008). Hence, further purpose of this letter is to
433 stimulate debate on the positive and negative aspects of rigid model structures or
434 component-based ones, and how the review process can best evaluate them.

435

436 **Reliability**

437 Model reliability cannot be assessed regardless of a presumed range of accuracy. A
438 specific parameterization of a model can be considered as reliable, if it produces results
439 that fall within a well-defined range of accuracy. In the case of a predictive model, the
440 range of accuracy can be defined statistically, proceeding from tests against observations.
441 In the case of an exploratory model, the range of accuracy may be defined through
442 sensitivity analysis, assuming that inaccuracy results from uncertainty in the values of
443 model parameters (e.g. Confalonieri et al., 2010).

444

445 Reliability is also a key aspect of credibility, where measures are influenced by the ability
446 to establish reliability with available past observations. It cannot be assumed, however,
447 that statistical (or any numerical) analysis is all that is required for model outputs to be
448 accepted particularly when models are used with and for stakeholders. The numerical
449 analysis provides credibility within the techno-scientific research community yet, while
450 necessary; this may be insufficient to achieve credibility with decision makers and other
451 stakeholders. Possibly a real test of model validity is whether stakeholders have sufficient
452 confidence in the model to use it as the basis for making management decisions (Vanclay
453 and Skovsgaard, 1997).

454

455 Reliability can also be interpreted as versatility of the model, that is, how well does the
456 model perform in situations for which it was not originally designed, or respond to
457 extreme conditions beyond that which calibration data represent? Sometimes, it is

characterized by a ratio of the real word observed data described by the model outputs (Mankin et al., 1975). The assessment of model versatility is based on the qualitative analysis of model structure (i.e. mathematical expressions) and potential results the model in question can generate. In many cases, only qualitative assessment can lead to subjective conclusions. The quantification of the concept is difficult or hardly possible due to limited observation data that is insufficient to understand environmental behaviour, model complexity limiting evaluation of possible model outcomes, and uncertainty in modelling results.

How to legitimate model usage

For well-developed environmental applications, model evaluation and selection techniques are heavily influential and can be used to build scientific and perceived credibility. However, establishing credibility is not straightforward for larger-scale environmental applications with many sources of uncertainty, decision-makers with different interests, and plausible future states that can be markedly different from observed past states. In these cases, credibility can be influenced by subjective measures and contingencies in the decision-making process (e.g., Aumann, 2008).

Establishing model credibility with end users / stakeholders can be problematic since they may have preconceived, and sometimes immovable, conceptions (Carberry et al., 2002). The task then falls on the model developers to show sufficient evidence in a form understandable by the end user to persuade them to challenge their beliefs and to consider alternatives.

481

482 Given the number of potential outcomes and stakeholders involved, more inclusive
483 modelling approaches such as multiple model and ensemble forecasting approaches can
484 be useful for establishing credibility. In particular, approaches that allow for multiple
485 model inference where differing models and perspectives are not excluded (e.g., Min and
486 Hense 2006). Instead, different models are weighted and synthesized using quantitative
487 criteria such as statistical support. This is important in determining quantitative reliability
488 and model evaluation (Burnham and Anderson, 2002), but can also assist with qualitative
489 aspects of credibility when models are used to inform a contentious decision process. The
490 resulting consideration of multiple models serves as a proxy for including different
491 scientific and subjective views of how environmental systems function and the resulting
492 ensemble forecasts are considered to be more broadly representative of the perspectives
493 of the decision-making participants.

494

495 Our view is that the key to successfully legitimating model usage is making model
496 outputs be seen by stakeholders as relevant to their decision making process. Legitimacy
497 of model usage can be seriously compromised when research outputs refer to geographic,
498 temporal, or organizational scales that do not match those of decision-making. Hence,
499 though adequate assessment and evaluation of a model in one location may be shown,
500 acceptance by stakeholders may be limited when applied where testing has not been
501 conducted.

502

Where models are used for decision support or evidence based reasoning, credibility is a complex mix of social, technological and mathematical aspects that require developers to include social networking (between developers, researchers and end users / stakeholders) to determine model rationale, aim, structure etc., and importantly a sense of co-ownership. Again, evidence of testing and results from a standardised peer-review procedure aids dialogue with stakeholders, as the researchers applying the models can demonstrate independent testing.

In this respect, a key component to credibility building is that a model should make available all the key management options that the decision maker considers important and should to an acceptable degree respond to management interventions in a way that matches with the decision maker's experience of the real system. In terms of models of natural processes, management can be substituted with alternatives, such as external shocks and/or perturbations to the drivers of the system.

Using the 'see-saw' analogy, where environmental models' estimates are used in contentious issues, credibility becomes the focal balance point around which opposing parties construct their arguments. Hence, credible models can serve to unite opposing parties, rather than serve to allow them to argue at increasing distance from each other's viewpoints and expertises. For such cases, subjective decisions on the selection and assessment of evidence may be as important as the accuracy of the measurement or forecasting of a particular phenomenon (Matthews et al., 2008).

Lack of transparency is frequently cited as the reason for the failure of model based approaches. It is important to challenge some of the assumptions and conclusions that are drawn on how to respond to the issue of transparency. One response is to make models simpler and hence the argument becomes easier to understand. Yet while simplicity is in itself desirable (Raupach and Finnigan, 1988) and the operation of simpler models may indeed be easier to understand, it may well be that the interpretation of their outputs is no simpler and indeed their simplicity may mean that they lack the capability to provide secondary data which can ease the process of interpretation. There is also a trade-off between simplicity and flexibility and this flexibility may be a crucial factor in allowing the tools to be relevant for counter-factual analyses. The current best practice for balancing simplicity and flexibility within the model development process seems to be the reusable component approach combined with a flexible model integration / evaluation environment. A set of standards applied within the peer review procedure therefore needs to address the issues of simple versus complex models and so look beyond the numerical testing and consider the flexibility of the model, ability of it to shed new light on an environmental issue, and aid the process of interpretation.

A constraint to both scientific credibility and transparency of models is the necessarily inherent inter-dependency of the modelled processes. A ‘fault’ in a model may be difficult to locate as many other related modelled processes confound it. Similarly, an effective modelled description of a specific sub-process may not be readily identified due to its dependence on less than satisfactory descriptions of other system features. Ranges of sensitivity and uncertainty analyses have been deployed to address this issue, although

the results are not always easy to interpret in terms of the original model formulation. Comparison of alternative model formulations can provide useful information in this context (e.g. Confalonieri et al., 2009a) but still suffers from the difficulty of disentangling inter-dependencies in the model. Crout et al. (2009) have proposed simple model reduction methods based on the approach of Cox et al (2006) which systematically explored the role of individual model variables on the models' predictive performance. This procedure frequently locates variables whose formulation has a detrimental effect on model performance.

The way forward: standardized evaluation tools

Based on the above views and highlighting of issues, we now explore options for future model evaluation. Turning back to the fact that model developers are normally lacking resources for adequate model evaluation, we conclude that introducing a more formalized set of evaluation criteria demands a standardised set of evaluation tools.

Identifying the prior art

It is common to believe that the total amount of environmental models is huge and that they cover almost all environmental situations. Nevertheless, the recent review of models used by the European Environment Agency in its recent environmental assessments and reports identified gaps in the availability, accessibility and applicability of current modelling tools (EEA, 2008). Indeed, journal articles reporting modelling efforts normally focus on the scientific interpretation of the findings, not on model documentation. There is no guarantee that a model, on which a published article several

years ago, is still readily available or even existing in any form that makes it possible to reproduce reported results. Can we therefore consider models that are not readily and completely available as the prior art? Scientific etiquette suggests that model documentations must be conveniently accessible, complete and mutually comparable (Benz and Knorrenschild, 1997; Voinov et al., 2009). We therefore suggest a register of models that can be considered as the prior art is needed for the technical assessment and evaluation of newly developed models.

Developing a comprehensive numerical library

A disciplined approach, effective management, and well-educated personnel are some of the key factors affecting the success of a software development project. Professionals in environmental modelling can learn a lot from software engineering, commercial product testing (especially in aircraft design and other areas where there is a very high safety standard required), stakeholders' deliberation and scientific developments from other disciplines. In so doing, we can expand our horizons to include the necessary knowledge to conduct successful model evaluation. Whilst some research has been undertaken focusing on establishing a baseline for evaluation practice, rather less work has been done to develop a basic, scientifically rigorous approach to be able to meet the technical challenges we currently face. We believe model evaluation software tools can valuably support this activity, allowing consolidated experience in evaluating models to be formed and shared. Whether model evaluation is a scheduled action in modelling projects, little work is published in the open literature (e.g., conference proceedings and journals) describing the evaluation experience accumulated by modelling teams (including

interactions with the stakeholders). Failing to disseminate the evaluation experience may result in the repetition of the same mistakes in future modelling projects. Based on past experience, establishing a better quality assurance program for a new modelling project may certainly increase the probability of success for that project. Learning from the experience of others is an excellent and cost-effective educational tool. The return on such an investment can easily be realized by preventing the failures of modelling projects and by avoiding wrong simulation-based decisions. Where complex models are to be evaluated, options are available to combine detailed numeric and statistical tests of components and sub-processes with a deliberative approach for overall model acceptance. Future model development should aim to incorporate automated evaluation checks using embedded software tools, with the aim of achieving greater cost and time efficiency and to achieve a higher level of credibility. Information from evaluation tools employed by the model developers needs to be made available to the peer review process. Beyond this, providing third parties with the capability of extending methodologies without re-compiling the component will ensure greater transparency and ease of maintenance, also providing functionalities such as the test of input data versus their definition prior to computing any simple or integrated evaluation metric. Making it in agreement with the most modern developments in software engineering, components for model evaluation will better serve as a convenient means to support collaborative model testing among the network of scientists involved in creating component-oriented models in the environmental field.

Moving from software to webware

Modern information and communication technologies offer the opportunity for a revolution in the area of technical assessments of environmental models (Alexandrov and Matsunaga, 2008; Hoffman et al., 2008). Moving from software to webware makes models (and data used to test them) available through a web-browser. It seems a time has come to think seriously about an Environmental Modelling Server (EMS) - a supercomputer (or a computing grid) for deploying environmental models and running them through web-browsers. The EMS may also do the routine work on technical assessment of models, providing the necessary resource currently lacking.

Our position

Concluding this letter, we emphasize that having standardized evaluation tools is the issue that needs to be tackled. Standardised model evaluation can consist of evaluation tools for use during and after the model development process, which can feed into a codified procedure during peer review of articles based on the model. The articles published in this thematic volume of EM&S are suggesting “*the evaluation of models should be a central part of the model development process, not an afterthought*” (Crout et al., 2009). This implies a clear demand for relevant software tools and acceptance by journals to adopt a minimum standard for peer review. Evaluation tools, in contrast to models, are generic by their nature, based on shared information and on re-using data from previous research exercises. The burden of developing evaluation tools is too hard for every single modeller. This and improving the peer review process are tasks that need

a communal effort based on International Environmental modelling and Software
Society's (iEMSs) leadership.

References

- Alexandrov, G.A., 2006. The purpose of peer review in the case of an open-access
publication. *Carbon Balance and Management*, 1:10
[<http://www.cbmjournal.com/content/1/1/10>]
- Alexandrov, G.A., Matsunaga, T., 2008. Evaluating consistency of biosphere models:
software tools for a web-based service. In: Sánchez-Marrè, M., Béjar, Comas, J.J.,
Rizzoli, A.E., Guariso, E., (Eds.), *Proceedings of the iEMSs Fourth Biennial Meeting:*
International Congress on Environmental Modelling and Software (iEMSs 2008).
International Environmental Modelling and Software Society, Barcelona, Catalonia.
[[http://www.iemss.org/iemss2008/uploads/Main/S12-02-Alexandrov_et_al-](http://www.iemss.org/iemss2008/uploads/Main/S12-02-Alexandrov_et_al-IEMSS2008.pdf)
[IEMSS2008.pdf](http://www.iemss.org/iemss2008/uploads/Main/S12-02-Alexandrov_et_al-IEMSS2008.pdf)]
- Aumann, C., 2008. A methodology for building Credible Models for Policy Evaluation.
In: Sánchez-Marrè, M., Béjar, Comas, J.J., Rizzoli, A.E., Guariso, E., (Eds.),
Proceedings of the iEMSs Fourth Biennial Meeting: International Congress on
Environmental Modelling and Software (iEMSs 2008). International Environmental
Modelling and Software Society, Barcelona, Catalonia.
[http://www.iemss.org/iemss2008/uploads/Main/S12-01-Aumann_et_al-IEMSS2008.pdf]

659

660 Bellocchi, G., Confalonieri, R., Donatelli, M., 2006. Crop modelling and validation:
661 integration of IRENE_DLL in the WARM environment. Italian Journal of
662 Agrometeorology 3, 35-39.

663

664 Bellocchi, G., Acutis, M., Fila, G., Donatelli, M., 2002. An indicator of solar radiation
665 model performance based on a fuzzy expert system. Agron. J. 94, 1222-1233.
666 [<http://agron.scijournals.org/cgi/reprint/94/6/1222.pdf>]

667

668 Bellocchi G., E. Habyarimana, M. Donatelli, M. Acutis, 2008. A software component for
669 model output evaluation. In: Sánchez-Marrè, M., Béjar, Comas, J.J., Rizzoli, A.E.,
670 Guariso, E., (Eds.), Proceedings of the iEMSs Fourth Biennial Meeting: International
671 Congress on Environmental Modelling and Software (iEMSs 2008). International
672 Environmental Modelling and Software Society, Barcelona, Catalonia.
673 [[http://www.iemss.org/iemss2008/uploads/Main/S12-06-Donatelli_et_al-](http://www.iemss.org/iemss2008/uploads/Main/S12-06-Donatelli_et_al-IEMSS2008.pdf)
674 IEMSS2008.pdf]

675

676 Bellocchi G., Rivington M., Donatelli M., Matthews K.B., 2010. Validation of
677 biophysical models: issues and methodologies. A review. Agron. Sustain. Dev., 30, 109-
678 130..

679

680 Benz, J., Knorrnschild, M., 1997. Call for a common model documentation etiquette.
681 Ecological Modelling 97, 141-143.

682

683 Beven, K., 2006. A manifesto for the equifinality thesis. *Journal of Hydrology* 320, 18–

684 36.

685

686 Beven K.J., Freer J., 2001. Equifinality, data assimilation, and uncertainty estimation

687 in mechanistic modelling of complex environmental systems using the GLUE

688 methodology. *J. Hydrol.* 249, 11-29.

689

690 Burnham, K.P., Anderson, D.R., 2002. Model selection and multimodel inference: a

691 practical information-theoretic approach, 2nd ed., Springer-Verlag: New York (NY).

692

693 Carberry, P.S., Hochman, Z., McCown, R.L., Dalglish, N.P. and others, 2002. The

694 FARMSCAPE approach to decision support: farmers', advisers, researchers' monitoring,

695 simulation, communication and performance evaluation. *Agricultural Systems* 74, 141–

696 177.

697

698 Cheng, R.T., Burau, J.R., Gartner, J.W., 1991. Interfacing data analysis and numerical

699 modelling for tidal hydrodynamic phenomena. In: Parker, B.B. (Ed.) *Tidal*

700 *hydrodynamics*, John Wiley & Sons: New York, pp. 201-219.

701

702 Confalonieri, R., Acutis, M., Bellocchi, G., Donatelli, M., 2009a. Multi-metric evaluation

703 of the models WARM, CropSyst, and WOFOST for rice. *Ecological Modelling* 220,

704 1395-1410.

705

706 Confalonieri, R., Bellocchi, G., Boschetti, M., Acutis, M., 2009b. Evaluation of
 707 parameterization strategies for rice modelling. Spanish Journal of Agricultural Research
 708 7, 680-686.

709

710 Confalonieri, R., Bellocchi, G., Tarantola, S., Acutis, M., Donatelli, M., Genovese, G.,
 711 2010. Sensitivity analysis of the rice model WARM in Europe: Exploring the effects of
 712 different locations, climates and methods of analysis on model sensitivity to crop
 713 parameters. Environmental Modelling & Software 25, 479-488.

714

715 Council of Science Editors, 2006. CSE's White Paper on Promoting Integrity in Scientific
 716 Journal Publications, CSE, Reston, Va.
 717 [[http://www.councilscienceeditors.org/editorial_policies/whitepaper/entire_whitepaper.p](http://www.councilscienceeditors.org/editorial_policies/whitepaper/entire_whitepaper.pdf)
 718 [df](http://www.councilscienceeditors.org/editorial_policies/whitepaper/entire_whitepaper.pdf)]

719

720 Cox, G.M., Gibbons, J.M., Wood, A.T.A., Craigon, J., Crout, N.M.J., 2006. Towards the
 721 systematic simplification of mechanistic models. Ecological Modelling 198, 240–246.

722

723 Crout, N., Kokkonen, T., Jakeman, A.J., Norton, J.P., Newham, L.T.H., Anderson, R.,
 724 Assaf, H., Croke, B.F.W., Gaber, N., Gibbons, J., Holzworth, D., Mysiak, J., Reichl, J.,
 725 Seppelt, R., Wagener, T., Whitfield, P., 2009. Good modelling practice. In: Jakeman,
 726 A.J., Voinov, A.A., Rizzoli, A.E., Chen, S.H. (Eds.) Environmental modelling, software
 727 and decision support. Elsevier, pp.15-31.

728

729 EEA, 2008. Modelling environmental change in Europe: towards a model inventory
730 (SEIS/Forward). EEA Technical Report 11/2008, European Environment Agency,
731 Copenhagen.

732

733 French, S., Geldermann, J., 2005. The varied contexts of environmental decision
734 problems and their implications for decision support. *Environ. Sci. Policy* 8, 378-391.

735

736 Funtowicz, S.O., Ravetz, J.R., 1990. *Uncertainty and Quality in Science for Policy*.
737 Kluwer Academic Press: Dordrecht (The Netherlands).

738

739 Gaber, N., Pascual, P., Stiber, N., Sunderland, E., Cope, B., Nold, A., 2008. Guidance on
740 the Development, Evaluation and Application of Environmental Models, Council for
741 Regulatory Environmental Modeling, U.S. Environmental Protection Agency,
742 Washington.

743

744 Gardner, R.H., Urban, D.L., 2003. Model validation and testing: Past lessons, present
745 concerns, future prospects. In: Canham, C.D., J.J. Cole, Lauenroth, W.K., (Eds.), *Models*
746 *in Ecosystem Science*. Princeton (NJ): Princeton University Press, pp. 184-203.

747

748 Glasow, P.A., Pace, D.K., 1999. SIMVAL '99: making VV&A effective and affordable
749 workshop, The Simulation Validation Workshop 1999, January 26-29, Laurel, MD, USA.

750

Haag, D., Kaupenjohann, M., 2001. Parameters, prediction, post-normal science and the precautionary principle - a roadmap for modelling for decision-making. *Ecological Modelling* 144, 45-60.

Hames, I., 2007. Peer-review and manuscript management in scientific journals: guidelines for good practice. Blackwell Publishing; Oxford.

Harremoës, P., 2003. The role of uncertainty in application of integrated urban water modeling. In: *Proceedings of the International IMUG Conference, Tilburg, 23–25 April 2003*.

Hoffman, F., G. Bonan, C. Covey, I. Fung, Y.-H. Lee, J. Randerson, S. Running, P. Thornton, The Carbon-Land Model Intercomparison Project (C-LAMP): A Protocol and Evaluation Metrics for Global Terrestrial Biogeochemistry Models. In: *Sánchez-Marrè, M., Béjar, Comas, J.J., Rizzoli, A.E., Guariso, E., (Eds.), Proceedings of the iEMSs Fourth Biennial Meeting: International Congress on Environmental Modelling and Software (iEMSs 2008). International Environmental Modelling and Software Society, Barcelona, Catalonia. [http://www.iemss.org/iemss2008/uploads/Main/S12-03-Hoffmann_et_al-IEMSS2008.pdf]*

Hsu, M.H., Kuo, A.Y., Kuo, J.T., Liu, W.C., 1999. Procedure to calibrate and verify numerical models of estuarine hydrodynamics. *J. Hydr. Engrg.* 125, 166-182.

774 Huth, N., Holzworth, D., 2005. Common sense in model testing. In: Zerger, A., Argent,
 775 R.M. (Eds.), Proceedings of MODSIM 2005 International Congress on Modelling and
 776 Simulation: Advances and applications for management and decision making, 12-15
 777 December, Melbourne, Australia, p.2804-2809.
 778
 779 Jakeman, A., Chen, S., Rizzoli, A., Voinov, A.A., 2009. Modelling and software as
 780 instruments for advancing sustainability. In: Jakeman, A.J., Voinov, A.A., Rizzoli, A.E.,
 781 Chen, S.H. (Eds.) Environmental Modelling, Software and Decision Support, Elsevier,
 782 pp. 345-366.
 783
 784 Jakeman, A.J., Letcher, R.A., Norton, J.P., 2006. Ten iterative steps in development and
 785 evaluation of environmental models. Environmental Modelling and Software 21 (5), 602–
 786 614.
 787
 788 Janssen, P.H.M., Heuberger, P.S.C., 1995. Calibration of process-oriented models. Ecol.
 789 Model. 83, 55-66.
 790
 791 Jørgensen, S.E., Fath, B.D., Grant, W., Nielsen, S.N., 2006. The editorial policy of
 792 Ecological Modelling. Ecological Modelling 199, 1-3.
 793
 794 Keating, B.A., Gaydon, D., Huth, N.I., Probert, M.E., Verburg, K., Smith, C.J., Bond,
 795 W., 2002. Use of modelling to explore the water balance of dryland farming systems in
 796 the Murray-Darling Basin, Australia. Eur. J. Agron. 18, 159-169.

797

798 Leffelaar, P.A., Meike, H., Smith, P., Wallach, D., 2003. Modelling cropping systems -
799 highlights of the symposium and preface to the special issues. 3. Session B. Model
800 parameterisation and testing. Eur. J. Agron. 18, 189-191.

801

802 Li, W., Arena, V.C, Sussman, N.B., Mazumdar, S. 2003. Model validation software for
803 classification models using repeated partitioning: MVREP. Comput. Meth Prog. Bio. 72,
804 81-87.

805

806 Mankin, J.B., O'Neill, R.V., Shugart, H.H. and Rust B.W. 1975. The importance of
807 validation in ecosystem analysis. In New directions in the analysis of ecological systems,
808 Part 1. La Jolla, CA: Society for computer simulation, v. 5: 63-71.

809

810 Marcus, A.H., Elias, R.W., 1998. Some useful statistical methods for model validation.
811 Environ. Health Persp. 106, 1541-1550.

812

813 Matthews, K.B., Rivington, M., Blackstock, K., Buchan, K., Miller, D.G., 2008. Raising
814 the bar - Is evaluating the outcomes of decision and information support tools a bridge
815 too far? In: Sánchez-Marrè, M., Béjar, Comas, J.J., Rizzoli, A.E., Guariso, E., (Eds.),
816 Integrating sciences and information technology for environmental assessment and
817 decision making. Proceedings of the 4th Biennial Meeting of the International
818 Environmental Modelling and Software Society, 7-10 July, Barcelona, Spain, Vol. 1,
819 p.948-955.

820

821 Min, S.-K., Hense, A., 2006. A Bayesian Assessment of Climate Change Using
822 Multimodel Ensembles. Part I: Global Mean Surface Temperature. *Journal of Climate*,
823 19:3237-3256.

824

825 Müller, U. T., 2009. Peer-Review-Verfahren zur Qualitätssicherung von Open-Access-
826 Zeitschriften – systematische Klassifikation und empirische Untersuchung, PhD thesis,
827 Humboldt-Universität zu Berlin, 2009; available at [http://edoc.hu-](http://edoc.hu-berlin.de/docviews/abstract.php?id=29636)
828 [berlin.de/docviews/abstract.php?id=29636](http://edoc.hu-berlin.de/docviews/abstract.php?id=29636)

829

830 Murray-Darling Basin Commission, 2000. Murray-Darling Basin Commission,
831 Groundwater Flow Modelling Guideline, Murray-Darling Basin Commission, Project no.
832 125. Canberra.

833

834 Pachepsky, Y.A., Guber, A.K., Van Genuchten, M.T., Nicholson, T.J., Cady, R.E.,
835 Simunek, J., Schaap, M.G., 2006. Model abstraction techniques for soil water flow and
836 transport, 175 pp, Nuclear Regulatory Commission, Washington, DC; NUREG/CR-6884.

837

838 Parker, V.T., 2001. Conceptual problems and scale limitations of defining ecological
839 communities: a critique of the CI concept (Community of Individuals). *Perspect. Plant*
840 *Ecol. Evol. Syst.* 4, 80-96.

841

842 Raupach, M.R., Finnigan, J.J., 1988. Single layer models of evaporation from plant
843 canopies are incorrect, but useful, whereas multilayer models are correct, but useless:
844 discussion. *Aust. J. Plant Physiol.* 15, 705-716.
845

846 Reichert, P., Borsuk, M.E., 2005. Does high forecast uncertainty preclude effective
847 decision support? *Environmental Modelling and Software*, 20: 991-1001.
848

849 Reynolds, J.F., Ford, E.D., 1999. Multi-criteria assessment of ecological process models.
850 *Ecology* 80, 538-553.
851

852 Robson, B., Hamilton, D., Webster, I., Chan, T., 2008. Ten steps applied to development
853 and evaluation of process-based biogeochemical models of estuaries. *Environmental*
854 *Modelling & Software* 23 (4), 369–384.
855

856 Rykiel, E.J. Jr., 1996. Testing ecological models: The meaning of validation. *Ecological*
857 *Modelling* 90 229–244.
858

859 Sinclair, T.R., Seligman, N., 2000. Criteria for publishing papers on crop modelling.
860 *Field Crops Res.* 68, 165-172.
861

862 STOWA/RIZA, 1999. STOWA/RIZA, Smooth Modelling in Water Management, Good
863 Modelling Practice Handbook STOWA Report 99–05, Dutch Department of Public

864 Works, Institute for Inland Water Management and Waste Water Treatment (1999) ISBN
 865 90-5773-056-1 Report 99.036.

866

867 Vanclay, J.K., 1994. Modelling forest growth and yield, CAB International, Wallingford,
 868 United Kingdom.

869

870 Vanclay, J.K., Skovsgaard, J.P., 1997. Evaluating forest growth models. *Ecol. Modell.*
 871 98, 1-12.

872

873 Van der Sluijs, J. P., 2007. Uncertainty and Precaution in Environmental Management:
 874 insights from the UPEM conference. *Environ Modelling and Software*, 22(5), 590-598.

875

876 Van Nes, E.H. and Scheffer, M., 2003. Alternative attractors may boost uncertainty and
 877 sensitivity in ecological models. *Ecological Modelling*, 159, 117–124.

878

879 Van Oijen, M., 2002. On the use of specific publication criteria for papers on process-
 880 based modelling in plant science. *Field Crops Res.* 74, 197-205.

881

882 Vischel, T., Pegram, G., Sinclair, S., Wagner, W., Bartsch, A., 2007. Comparison of soil
 883 moisture fields estimated by catchment modelling and remote sensing: a case study in
 884 South Africa. *Hydrol. Earth Syst. Sci. Discuss.* 4, 2273–2306.

885

886 Voinov, A., Hood, R.R., Daues, J.D., Assaf, H., Stewart, R., 2009. Building a community
 887 modelling and information sharing culture. In: Jakeman, A.J., Voinov, A.A., Rizzoli,
 888 A.E., Chen, S.H. (Eds.) Environmental Modelling, Software and Decision Support,
 889 Elsevier, pp. 345-366.
 890
 891 Wager, E., 2006. Ethics: What is it for? Nature: Web Debate – Peer-review.
 892 [<http://www.nature.com/nature/peerreview/debate/nature04990.html>]
 893
 894 Walker, T.J., 1998. Free Internet Access to Traditional Journals. American Scientist,
 895 86:463
 896
 897 Wang, Y.-P., Trudinger, C.M., Enting, I.G., 2009. A review of applications of model-data
 898 fusion to studies of terrestrial carbon fluxes at different scales. Agricultural and Forest
 899 Meteorology, 149, 1829-1842.
 900
 901 Welsh, W., 2008. Water balance modelling in Bowen, Queensland, and the ten iterative
 902 steps in model development and evaluation. Environmental Modelling & Software 23 (2),
 903 195–205.
 904
 905 Ye, M., Meyer, P.D., Neuman, S.P. ,2008. On model selection criteria in multimodel
 906 analysis, Water Resources Research 44, W03428.
 907
 908

909

910

911

912

913

914 Figure 1. A flow-chart of peer-review process

915

916

Figure 2. A scheme for specifying a published model.