May 12, 2001

Appendix A

BAY-DELTA MODELING FORUM - GUIDELINES FOR A PEER-REVIEW PROCESS

BAY-DELTA MODELING FORUM
GUIDELINES FOR A PEER-REVIEW PROCESS
Submitted by the Bay-Delta Modeling Forum Peer Review Committee
to the Steering Committee
Version 1.0
Approved by the Steering Committee
September 10, 1996

Introduction

One of the most important goals of the Bay-Delta Modeling Forum (the Forum) is to provide a mechanism for evaluating models used in managing the Bay-Delta system. This task has been delegated to the Peer Review Committee (Committee). The Committee has taken as its first task the design and execution of a test-case peer review of one-dimensional hydrodynamic models of the Delta. This choice was made for a variety of reasons including the availability of funding specifically for such a review, the variety of potentially competing models of this class, and the high visibility of these models in the Bay-Delta scientific and engineering community.

This document maps out a proposed process by which the peer review may be completed. The first sections discuss the principles and practices that we expect to apply to nearly all peer reviews. The final section contains details about the proposed process as it

applies to the chosen test case. This proposal is presented to the Steering Committee for discussion, amendment as necessary, and approval.

Vision Statement

The peer review process is a mechanism by which reviews of models (or other analytical tools) can take place in a timely, open, fair, and helpful manner. The peer review serves two principal clients: model developers and model users. It does so by 1) providing constructive feedback to model developers, and 2) serving to further the models' acceptance and understanding by the user community through building confidence, reinforcing models' limitations, and providing guidance on model utility for a variety of tasks. All steps of the review process are open to the developer and user community and are intended to be continually improved.

General Principles

The review process is not intended to provide some stamp of approval for a particular model, nor to disapprove of a model. Rather, it is intended to illuminate the topics for which a given model is a suitable tool, and the temporal, geographic, or other limits on use of the model. This peer review process will not be available to review single models intended for limited use on specific applications, nor for critiquing the results of a model application in a specific use.

Requests for review may originate from either developers or users, or may be generated internally by the Committee. Although single models may be reviewed, when multiple models in the same functional class are available, the Committee will attempt to include them in a simultaneous, comparative review. Similarly, when models serve multiple purposes, those purposes will be individually reviewed. Data sets (or design strategies for collecting, distributing, or displaying data) intended to serve multiple users will also be candidates for review.

The review team is selected by the Committee and reports to the Committee. The Committee has the task of facilitating the review,

the workshops, and the process by which comments on the draft review are solicited and presented to the review team. In addition, the Committee is to work with the Executive Director to disseminate results of the review. The Committee will keep the Steering Committee apprised of all of its activities, and solicit assistance with funding issues from the Steering Committee.

Development of a peer-review process has been slowed by disagreement over several key issues. These are discussed in the following paragraphs along with the resolution arrived at by the Committee.

Membership of review team: Recommendations on who should sit on the review team have ranged from all model developers to all outsiders. The former provides the necessary degree of familiarity with the models, while the latter provides the necessary impartiality. Clearly the Solomonic solution is best: we should draw on both pools of talent with their characteristics. Thus, review teams should consist of agency personnel with at least two experts from other geographic areas. This should shorten the learning curve of the outside reviewers while reining in any partisan tendencies on the part of the agency people.

Need for continual review: Modeling is an ongoing process and, as models develop, review results become obsolete. This means that there is a danger of reviews affecting the use of models well beyond the time when the models have been revised. This is a gray area requiring careful attention by the Forum. We do not want to spend unnecessary effort on repeating reviews of a single model or class of models every time a small change is made. Resolution of this issue should wait at least until we have completed our first review, and have had a chance to evaluate both the review and its effect on the use and development of the models reviewed. In general, peer review of models is seen as a long-term activity of the Forum.

Participation by model developers: Models will not generally be reviewed except by sponsorship of the owners or developers. An exception may be made for models in the public domain, in which case the model users would be asked to sponsor the model. Initially a disincentive may exist for an owner to have a model reviewed, both because of the additional work required to prepare for the review and because the review exposes the model to criticism. The Committee hopes that this will eventually be offset by the perception among

regulators that peer-reviewed models are preferable to those that have not been reviewed.

Definitions

A "class" of models are models of a similar type that perform a similar function. For instance, 1-D numerical models of the Delta would be considered a single class of models. Multi-dimensional models of the Delta would be in a different class, because they have quite different uses and data requirements than 1-D models; black-box or statistical models of the Delta would be in yet another class, because they are not derived from fundamental equations of physics.

A "category" within a class is a type of function or output that most or all models in the class would have. For instance, in 1-D numerical models, different categories could be hydrodynamics, mass tracking, and particle tracking modules.

Peer Review Process

The peer-review process is presented below as a work-in-progress. No doubt elements of it will be found unworkable or needing improvement.

The steps of the process are:

1. Select the models: Models to be reviewed are selected by the originators of the review, whether they be the Committee, users, or developers of models to be reviewed.
2. Select Reviewers: The Committee accepts input from other Forum members and the originators of the review, then makes an initial selection of the review team. In general this team will comprise technical personnel familiar with the models to be reviewed, and outside and presumably unbiased experts in the field of the models being reviewed. The qualifications of the reviewers and, in particular, their experience with the models under review, will be clearly described in an appendix to the final report.
3. Obtain funding: The Committee works with the Steering Committee to identify sources of funding for the review. This may include the originating agencies, funds earmarked for peer review, or other sources.
4. Assemble Model, Documentation, and Data: In this stage, models

are assembled with documentation and needed data and provided to the review team along with the necessary documentation and data. This would be done by the model developers or sponsoring agencies at the request of the Committee.

5. Scope the review: The Committee works with the review team, the Steering Committee, model developers, and other interested parties to develop a specific plan of action for the review. This plan will include the specific objectives of the review, a detailed set of guidelines for the reviewers, and a schedule.

6. Conduct Initial Review: Members of the review team examine documentation and other materials provided to them, including the actual model if it is sufficiently portable. They then provide written questions to model developers, who may then respond. One or more workshops would be set up to review initial results anddevelop a collaborative discussion on the models. Participation in these workshops would be by invitation of the Committee, although requests by individuals or agencies to participate would normally be honored. The purpose of the workshops is to ensure a clear understanding of the models and achieve consensus on the capabilities and limitations of the models as indicated by the materials provided. In addition, a description of one or more proposed test runs would be provided to the Committee.

7. Test Models: Either the model review team or, in the case of complex or platform-specific models, the sponsoring organization, would perform the requested model runs. These would be summarized in a report of the model tests including parameter values used, conditions of the test, and results.

8. Prepare Draft Report: After all models have been reviewed, and obvious errors in the models or the manner in which they are used have been identified, the review team will develop a draft report. This report includes results of the previous 2 steps. This draft will be reviewed by the Committee before the next two steps are taken.

9. Conduct review workshops: Results of the review are presented to developers, users, sponsoring organizations, the Committee, and other interested Forum members. Comments are tabulated by a Committee member or designee, and provided to the review team.

10. Prepare Final Report: The review team revises its report as it sees fit, and includes in appendices the comments on the draft review, responses to comments, and any dissenting views.

Example Guidelines

Guidelines provided to the review team would vary with the model(s) being reviewed and the specific objectives of the review. The guidelines presented below are for example only to illustrate some of the considerations to be taken in reviewing the models. In this case it is assumed that several models of a particular class are being reviewed; the situation for one model would differ in ways that are apparent on reading this material. Also, some of these guidelines would be inapplicable to some kinds of models.

General

1. The review should proceed by consensus, not by democratic practices. Any disagreement that cannot be resolved is to be included in the final report as a comment or dissenting view.
2. All members of the review team should sign the final review document.
3. Take a non- confrontational attitude; the task is not to attack a particular model, but to convince developers and users of the value and accuracy of the review.
4. Pinpoint the good as well as the bad; do not hold the models to an impossible ideal.
5. Suggest areas for improvement of the particular modeling enterprise or field as a whole
6. Keep all comments as succinct as possible.
7. Live up to the imposed deadline.

Modeling Context

1. Define and explain the particular model class. Explain the limitations of this model class; what can this type of model examine, what can it not? Provide examples to illustrate positive and negative usage of models in this class.
2. Are there any fundamental differences among the questions each of these models was designed to address?
3. What are the major constraints that limit the domain of these models, i.e., under what circumstances would they be scientifically indefensible and how would the user know?

General Content of Reviews

1. Provide an overview, evaluating and comparing each of the models' fundamental components, e.g., transport, diffusion, chemical kinetics, mortality rates, reproduction, animal behavior, etc. Do not judge the models at this point.
2. What are the defining spatial and temporal scales for the models?
3. How much do the models depend on an external description of how man has or is influencing the system, e.g., operational decision criteria for controlling pumps?

Theoretical Basis

1. Are the fundamental components primarily developed from sound first principles, or are they largely based on empirical measurements? Is there a reason to prefer one or the other?
2. Contrast the models' algorithms, their derivations, and solution techniques. Do the algorithms adequately reflect the concepts being addressed? Are any "gimmicks" employed?
3. Identify areas for improvement, e.g., are major components oversimplified, missing, or too restrictive?
4. Are there any features of this model that raise enough concern to alert current users to their "faults?" What elements are marginal and need immediate attention?
5. What extra features are needed to model the prototype, e.g., channel barriers, or effects of land-based activity. The importance of each feature, and how well it is simulated

Parameter Estimation

1. What methods and data were used for parameter estimation and were these techniques adequate?
2. What domain or boundary conditions are implied by the parameterization?
3. Advantages and disadvantages of the discretization used, both spatial and temporal. How much can the user change discretization parameters (e.g., delta t, delta x, ...). When does the model exhibit 'poor' behavior at extreme values, and in what manner?

Data Requirements

1. What kinds of data are required to apply the models? Are these data readily available, or costly to obtain? Are data collection procedures repeatable?
2. How are missing data handled?
3. What kind of quality assurance must one have for the data? Are the models sensitive to data input errors?
4. Are there any special data elements required by this model not generally shared by others in its category/class?
5. For non-historical inputs, is provision made to supply routines that will generate those inputs?
6. Does the model recognize questionable data in the input and notify the user?

Key Assumptions

1. What are the key assumptions and how well have the authors outlined them?
2. How do the assumptions limit the domain of applicability of the models?
3. How sensitive or robust are the models to these limiting assumptions?

Verification, Calibration, and Validation

1. What criteria have been used to verify the model, i.e. to assess the accuracy with which the model code does what it is designed to do, and to evaluate the models' performance?
2. What type of study was the model calibrated for (e.g., intra-tidal? 20 year?) What portion of the model's domain was not calibrated (e.g., a model can handle flood flows, but was not calibrated in that regime). What data went into the calibration, and validation? What model parameters were adjusted, using what methodology or algorithm? Validation results; how 'well' does the model validate? Does the model show trends or predictability in its error plots with respect to either time, space, or other parameters?
3. Have the assessments been well-conceived and thorough? Have extremes been tested? Have parts been tested even if the whole

cannot be? Have tests covered long periods at multiple locations under a diversity of conditions?
4. What are the implications for the models' accuracy and precision?
5. Are there any evident biases in the models' predictions through space or time?
6. How do the models address uncertainty?
7. What additional data or validation (if any) is needed to improve the models' scientific basis?
8. How does the model respond in a sensitivity analysis? Does it seem unusually sensitive or non-sensitive to changes in inputs? If so, what is the probable cause?

Usability

1. Are the sets of documentation (user's manual, programmer's manual, test applications, solution methods, guidance documents) clear, comprehensive, and usable? Do they contain good references? Is there a clear history of changes (fixes and updates) that have been made?
2. Are the programs easy to use? How long would it take a technically literate user to learn to use this model? Is the input logically organized? Is the output given in formats useful for interpretability, and to feed into subsequent analyses or given to other audiences?
3. Are there provisions for catching input errors and preventing use of the model outside its intended domain? If the model is used outside its domain, does it fail gracefully (warning messages) or simply crash?
4. Are the test and sample input/output data sets useful in demonstrating the full range of the model?
5. What is the overall cost of an application?
6. Is technical assistance available? Training classes available? Special disciplines required? Is program development ongoing and disruptive, or stable?
7. Are the programs in the public domain?
8. Do the models provide the user any information to alert them that they are being used incorrectly?
9. Do the models provide explanatory power, i.e., does their use further understanding of the way the systems behave?
10. How well are the theory and usage of the model documented? How well is the code itself written and documented; would it be easy to

modify the code? Will the developer be available for assistance, and how fast will assistance be provided, by what medium (written, telephone, e- mail, ...).
11. How long will each model run on a standard problem on a standard platform?

Retrospective

1. How well do the models satisfy their intended purpose? Is there a pragmatic match between models and their intended users? Is there a better way?
2. Are there any red flags to (continued) use for regulatory or scientific purposes? (But don't throw out the baby with the bathwater.) Would you use it?
3. What is the expected lifetime of the version of the models you evaluated? Do the developers have a stated commitment to continued model improvements, to the state of the art? If so, do they have a demonstrated track record of responding to user needs?
4. What key research is necessary to refine or improve the model and/or the data bases on which they rely?
5. Is it likely that any of these models have been misused?
6. Any other helpful suggestions for model developers or users? Don't be afraid to be constructively blunt.
7. Were there any facets of this evaluation that the group did not feel confident addressing? Any minority reports?
8. What recommendations would you have to improve the peer review process for the next cycle?