Integrated Modeling Support
Delta Stewardship Council Contract #17400

Memo 4. Recommendations for Modeling Best Practices

January 2020

Prepared for:

Delta Stewardship Council
Ben Geske, P.E., Project Manager

Project Team:

Tetra Tech, Inc.
Bachand & Associates
Cramer Fish Sciences
HydroFocus, Inc.
Pax Environmental
University of California at Davis
University of California at Merced

Acknowledgements

This work was performed under contract (#17400) between the Delta Stewardship Council (DSC) and Tetra Tech Inc. Mr. Ben Geske of the Delta Stewardship Council served as the project manager. The DSC's Integrated Modeling Steering Committee, consisting of modeling experts from state and federal agencies and private consulting firms, provided guidance on the study.

Project Team

Tetra Tech: Sujoy Roy, Paul Hutton, Katherine Heidel, Tom Grieb, and Ali Tasdighi

Bachand Associates: Philip Bachand and Yan Liang

Cramer Fish Sciences: Travis Hinkelman

Hydrofocus: Steve Deverel

Pax Environmental: Tom Lagerquist

University of California at Merced: Josué Medellín-Azuara and Anna Rallings

University of California at Davis: Anne Visser and Leslie Panyanouvong

Contents

Figures

Tables

Executive Summary

The use of models is commonplace and is becoming increasingly important for assessing most societally important environmental problems. This document provides recommendations for best practices that are expected to enhance the utility of modeling efforts to decision-makers, stakeholders and to the modeling community in general. The guidance offered in this document applies equally to individual discipline-specific models and integrated models that combine knowledge from different disciplines. Many of these concepts are general and can also be considered by non-modelers who need to understand the scope of a modeling exercise at its inception and to make a judgement as to the utility of the results upon its completion. This document is focused on applications in San Francisco Bay, the Central Valley and Sacramento-San Joaquin Delta that pertain to areas of interest of the Delta Stewardship Council (in this report we often refer to these regions in shorthand form as the "Delta"); however, the recommendations for best practice are applicable outside of this geographic domain.

This work provides a summary of actions that need to be undertaken to improve the robustness of virtually all modeling exercises, including efforts that are relatively modest in scope. These actions include: defining modeling purpose; developing conceptual models to provide a compact and transparent representation of key processes to communicate with stakeholders and other technical specialists and to aid in model selection or development; preparing standardized datasets that can be used to replicate a modeling study and compare across models; verifying code to ensure that the theoretical framework has been correctly implemented; documenting the model calibration process; and evaluating model performance over new data sets. This work also recommends a broader exploration of model structure and bias, going beyond routine calibration and evaluation/validation exercises, especially when observed data do not adequately match model predictions. Finally, this work recommends that adequate documentation be developed and made readily available to meet the needs of users as well as current and future modelers.

Additional practices can be adopted to improve modeling, but imposing requirements for such actions may not be practical for all studies. Additional actions for major modeling studies (i.e. those studies tied to large societally consequential decisions) are identified separately. These additional actions include: peer review of model studies at various stages of implementation; model sensitivity analysis to identify key drivers; model uncertainty analysis; consideration of novel approaches to meet sensitivity and uncertainty analysis needs of complex models and model frameworks; consideration of alternative models for model studies (where available); post-audits (i.e., review and evaluation of historical model predictions in light of new field observations); and development and compatibility with exchange standards to enable data sharing across models.

The technical strength of a model can be established through the above steps. Nonetheless, there remain several non-technical issues that should be addressed to meet the broader goals of a modeling exercise. These non-technical issues include: development of a communication strategy for a modeling study; consideration of bias in many aspects of the model formulation; presentation of results across many audiences; building trust across the community that will be using the model results; overall user-friendliness of the modeling framework; and practices for sustaining the usefulness of a model over a long-term horizon.

To encourage adoption of the best practices identified in this work, we provide three summary sheets, corresponding to different stages of modeling. The purpose of the first sheet, designed as a checklist to be employed at inception of a modeling effort, is to enable various participants to agree on the basic features of the work to be done. The purpose of the second sheet is to evaluate and score a modeling exercise upon completion. The final sheet is to assess the overall life cycle of a modeling framework.

Glossary

Term	Definition
All-at-a-time (AAT)	A sensitivity analysis approach where all parameters can be varied at each iteration. Typically used with global sensitivity analysis.
Boundary condition	A condition that is required to be satisfied at all or part of the boundary of a region in which a set of differential equations is to be solved.
Calibration	The process of changing values of model parameters in a quantitative model to match or “fit” the model to field observations.
Code	Representation of the theoretical formulation of a model in computer language that serves as the basis for developing an executable model. In many cases, even for public-domain models, the underlying codes are not in the public domain.
Code verification	The process of testing the accuracy of the model’s computer representation of the theoretical formulation. This process includes code examination, testing bounding cases, and comparison against analytical solutions of underlying equations (when available).
Conceptual model	A high-level representation of inputs, interacting processes and drivers, and outputs for any kind of process (e.g., physical, biological, economic, etc.). Although a conceptual model may include quantitative information, it is often presented in non-quantitative form and serves to communicate the model structure in a transparent manner. A conceptual model may be developed as a communication tool following the completion of a modeling study, or, during the initiation of the project, the conceptual model serves as the basis for selection of or development of a quantitative model.
Domain	In this work, a specialized field of study.
Empirical/statistical model	A mathematical formulation of inputs and outputs with limited process representation; model parameters calibrated with observed data.

Term	Definition
Emulator	Computationally simplified model representations that use relationships between inputs and outputs. Emulators are typically developed to reduce the computational cost of model exploration.
Evaluation	A general term for a sequence of steps taken to understand the performance of a model following calibration. Evaluation may include comparison against independent input and output data sets, sensitivity analysis for key parameters, or uncertainty analysis.
Global Sensitivity Analysis (GSA)	A sensitivity analysis approach that analyzes the variability of model responses across the full parameter space.
Initial condition	The solution of a differential equation over time requires the definition of values at the inception of the solution, termed the initial conditions. Other types of formulations, such as time series models, may also need the definition of initial conditions.
Local Sensitivity Analysis (LSA)	A sensitivity analysis approach that analyzes model responses around a well-defined region of interest in the input parameter space.
Lumped model	A model that aggregates variable information over time and space for simplification, or because of limited data availability. In contrast, a distributed model may have greater spatial and temporal resolution.
Metadata	A set of data that describes and gives information about other data.
Model configuration	The process of specifying background characteristics for a model simulation, e.g. the physical representation of a water body. Model configuration is performed once the theoretical framework of a model has been developed and implemented.
Model framework	A general term for the theoretical implementation of a process-oriented model. A model framework will usually need to be configured for application to a specific geographic setting. Many models in common use are general purpose frameworks that can be configured to represent the same set of processes in different regions (for example, watershed models), whereas others are developed from the ground up as applicable to a single location, and the configuration is embedded within the general setup.
Model lifecycle	A term referring to the entire timeframe from conceptualization of a mathematical model to implementation in computer code, and to multiple cycles of application, revision, and reuse in one or many different domains. Major models generally require large investments and a lifecycle of many decades.
Model structure	The representation of model inputs, key processes and interactions, and outputs. A conceptual model may graphically communicate the model structure, but even where a conceptual model is not published, all process-based models require an underlying model structure. In the case of data-driven models, internal processes are generally not represented, and model structure refers to the inputs that are selected a priori to influence the outputs.
Model training	Similar to calibration and parameter estimation, but typically used in the context of machine learning. The process of adjusting empirical model constants to match model outputs and field observations. In the context of machine learning, the model constants may have no physical meaning.

Term	Definition
Monte Carlo simulation	A general solution approach in modeling analysis where key values (for example, parameter values in a model) are sampled randomly over a defined space to provide a range of conditions for testing.
Numerical model	Many quantitative models are represented by differential equations that cannot be solved exactly (i.e. analytically) because of domain or mathematical complexity. Numerical solutions (such as finite elements or finite differences) are commonly-used approaches to estimate the solutions of differential equations. Models that employ such numerical solutions are particularly common in the representation of physical and chemical processes, and are termed numerical models.
One-at-a-time (OAT)	A sensitivity analysis approach where one parameter is changed at each iteration. Commonly used with local sensitivity analysis.
Parameter estimation	Similar to calibration. The process of adjusting parameter values in a model such that the model output matches field observations within an acceptable error range.
Parameters	Parameters in a quantitative model represent numeric constants associated with key processes. Typically, these processes represent a feature of a natural system (for example, reaction rates or hydraulic conductivities), and may be known within a range. The process of parameter estimation is to find values that enable the model to fit observed data within an acceptable range.
Sensitivity analysis	The process of adjusting model parameters or inputs within a realistic range to explore the effect on, or sensitivity of, model outputs. Model sensitivity in a multi-parameter model may depend on the states of other parameters, and individual model outputs may be more or less sensitive to different parameters. A common goal of sensitivity analysis is to identify parameter(s) that have the greatest impact on key model outputs.
Statistical model	See “Empirical/statistical model” above.
Uncertainty analysis	Model inputs or parameter values are presented in a probabilistic form (i.e., as a distribution of values) to a calibrated model, and the effects on model output evaluated. Given that inputs and model parameters are known with different degrees of error, the goal of uncertainty analysis is to quantify the range of outputs in a modeling study.
Validation	A term in common use in many modeling communities, validation refers to the process of applying a calibrated model to an independent set of observed data to assess whether the model fit is acceptable. A criticism of the term validation is that the process does not prove that a model is valid, but rather demonstrates performance over a limited range of conditions. The term evaluation is sometimes recommended as an alternative.

1 Introduction

Models are commonly used for assessing most societally important environmental problems. Models are best thought of as tools for integrating data, exploring processes in a structured manner, and evaluating responses under historical conditions or projected future scenarios. Although full understanding of a system is rarely encapsulated in a model, these activities cannot be performed as efficiently without models; hence, their extensive use in many modeling fields or domains, including the environmental domain.

In this work, the term "environmental model" refers to analysis tools (typically quantitative) that are used to represent the behavior of physical, chemical, biological, economic, and social systems. These various system domains often interact; thus, an individual environmental model may encompass more than just one system. Although economic and social systems have varied modeling frameworks, for these domains our focus is on models where the natural environment is a driver. In the Delta, the use of environmental models is widespread in the representation of physical, chemical, and biological systems, and the use of these models continues to grow, especially in the domains of economic and social systems.

Models in general, and environmental models in particular, must strike a balance between the competing needs of accessibility and comprehensiveness (Figure 1). A model formulation that is more readily understandable or accessible may focus on key processes and provide a more simplified system representation while omitting more complex relevant drivers. A more comprehensive model formulation may represent many drivers and capture system complexity at the expense of greater challenges to implement, test, and explain.

Model developers have flexibility in how they choose to represent a system but are usually limited by one or more of the following constraints: availability of observed data, availability of time and human resources, and computational resource requirements.

Model development is a creative process that seeks to find the "right" or "best" course of action given the above constraints. However, given that an *a priori* "right" system representation rarely (if ever) exists, there is considerable subjectivity in the selection of modeling approach. For these reasons, there is no obvious way to know if the modeled representation of a problem is correct and credible, and additional testing must be performed to assess these features.

This document provides recommendations for best practices that are expected to enhance the utility of modeling efforts to decision-makers, stakeholders, and the modeling community in general. To provide context for the best practices, this chapter first characterizes modeling processes being applied in the Delta. This work is part of a larger study evaluating the current state and future opportunities for integrated modeling in the Delta (See Memo 2, *A Survey of Recent Integrated Modeling Applications in the Delta and Central Valley* and Memo 3, *Institutional & Technological Challenges and Solutions for Model Integration and Data*). The guidance proposed in this document applies equally well to individual discipline-specific models as well as integrated models that combine knowledge from different disciplines.

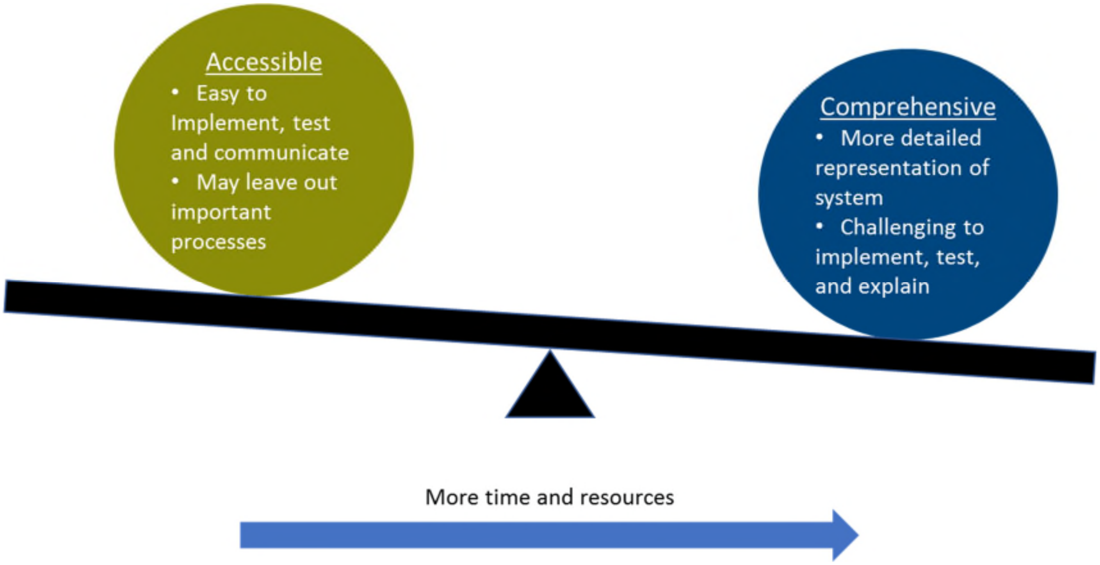

Accessible
• Easy to Implement, test and communicate
• May leave out important processes
Comprehensive
• More detailed representation of system
• Challenging to implement, test, and explain
More time and resources

Figure 1. Balance between accessible and comprehensive models.

1.1 Typical Uses of Models in the Delta

Commonly-used models in the Delta, as summarized in the *Model Inventory* (Memo 1), support a range of activities that can be broadly classified as follows:

- Planning and decision support - including but not necessarily limited to support for the development of new environmental regulations (e.g. changes to water quality standards), support for long-term facility or operational modifications (e.g. changes to reservoir operating rules), or support for the creation of new infrastructure (e.g. new alternatives for Delta conveyance or evaluation of new dam sites).
- Science support - including the generation and testing of hypotheses to better understand the Delta ecosystem. Activities include, for example, understanding the

population behavior of key species; understanding food web interactions; or understanding changes in landscape over the long-term due to human pressures, climatic change, and extreme events.

- Real-time operations support - including reservoir outflows for flood management and water supply, water exports from the Delta, and barrier operations used to manage salinity at various locations.
- Dispute settlement support - including legal proceedings in the context of water rights adjudication or allocation of water among different types of uses.

Given the growing pressures on water resources in California and the greater awareness of environmental protection, these activities will continue to need the support of credible modeling well into the foreseeable future.

1.2 Individual and Institutional Roles in Modeling

With the exception of science-based modeling, a variety of participants are involved in directing, executing, and evaluating the outcomes of a modeling study as shown in Figure 2. These participants may belong to different institutions or organizations with different areas of interest and expertise. Appreciating the intricacies of this typical structure is a prerequisite to understanding how modeling best practices can effectively serve these varied participants. Thus, a model study will have a **sponsor**, which is an institution or group of institutions that have an interest in the outcome and provide the resources for its performance. The sponsor will likely broadly define the scope of the model study, including question(s) to explore, scenarios of interest, schedule, funding, etc. The actual development, testing, and reporting of a model study will likely be performed by **model specialists** with knowledge of the specific domain and with relevant software development skills. In many cases, the sponsor and other decision makers will not work directly with model specialists or model results. Rather, one or more **domain experts** may help with interpreting and communicating model results to the sponsor. Finally, **stakeholders** with an interest in the outcome of a study may influence the process through the model sponsor or the decision makers. In some instances, stakeholders may directly interact with model specialists. Indeed, with the growing application of models in many areas of decision-making, it is desirable to engage and enable stakeholders to play a larger role in modeling studies, in the Delta and globally (Voinov and Bousquet, 2010; Voinov et al., 2016).

Models that are focused on scientific advancement and led by research teams (the manner in which most scientific research is conducted in the U.S.) may have a simpler structure than shown in Figure 2. Although such models may not be directly used in policy-making in their early phases, they serve two roles: (i) the models may mature over time and drive larger scale policy decisions, as described in the next section, or (ii) the new understanding and related individual expertise gradually diffuses into the broader modeling community.

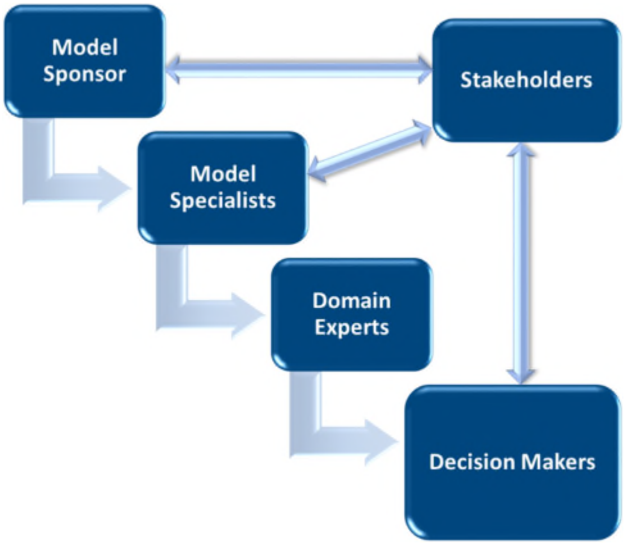


Figure 2. Key roles in modeling studies.

1.3 Model Types

Several different mathematical approaches may be applied in the development of quantitative environmental models as shown in Table 1. The broad classes of mathematical approaches in use include analytical/numerical solution of process equations over a defined domain; statistical/empirical models that are based on relationships between observed data but typically contain little to no process representation; optimization based models that seek to meet key objectives subject to a set of defined constraints; machine learning based models, a sub-class of statistical/empirical models with a wider range of algorithms and capacity to handle disparate data sets; and agent-based models that represent behavior of organisms or populations (animal or human) in response to external factors. Several of these approaches may be combined within a single modeling system, resulting in a “hybrid” model. As described in the following chapters, the underlying approach adopted within a particular modeling framework affects the applicable best practices for development.

Table 1. Types of Models used for Environmental Modeling

Model type	Feature
Analytical/Numerical	Solving a framework of process equations, either in closed analytical form or numerically; model parameters calibrated with observed data.
Statistical/Empirical	Limited process representation; model parameters calibrated with observed data.
Optimization based	Focused on meeting key objectives under a range of input conditions.
Machine-learning based	Trained on finding patterns or relationships in available data, but with minimal process-oriented representation. These are an extension of the statistical/empirical models, but with a greater variety of emerging algorithms to represent increasingly complex data sets.
Agent-based	Represents behavior of organisms or populations (animal or human) in response to external factors over time and space.

1.4 Model Elements

For most of the model types described above, the system is composed of the elements shown in Figure 3. The model is driven by initial and boundary conditions of the variables of interest, where the initial conditions represent the values at the beginning of the model run and the boundary conditions represent values at the edge of the domain to be modeled. Specification of initial and boundary values influence the time evolution and spatial scale of model calculations. The model configuration is used to define the background setting over which the calculation is being performed, such as the bathymetry of a water body or the depth of an aquifer. Within the model, there are usually some pre-defined or adjustable parameters. Pre-defined parameters refer to values that are independently measured or known, such as the properties of water density as a function of temperature. Adjustable parameters are typically those that cannot be measured directly and are derived by fitting the model to observed data (a process called calibration which is described in the following chapter). The model may calculate values (over time or space) based on the equations, configuration, and boundary conditions, termed the internal state variables. A subset of or an interpreted summary of the state variables may be presented as outputs; outputs may be presented in tabular form or in various graphical forms. Best practices for modeling are related to each of these elements.

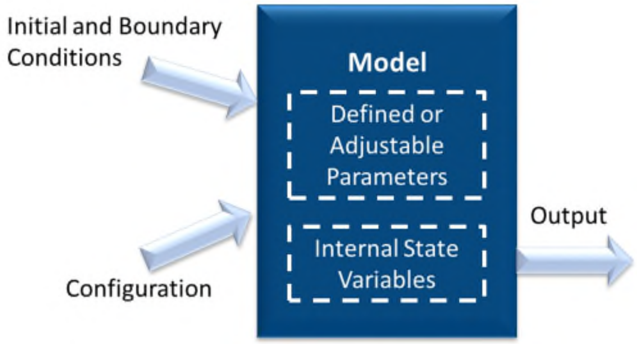

Initial and Boundary Conditions
Model
Defined or Adjustable Parameters
Output
Internal State Variables
Configuration

Figure 3. Major elements in model systems.

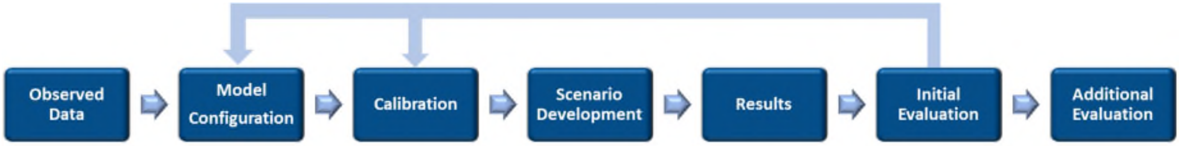
1.5 Model Process

The steps necessary to model a specific problem depend on the nature and history of the problem being studied. When basic principles of the problem are well understood and mature, a model study will likely utilize an existing model framework, customized for a specific location of interest. When basic science associated with the problem is still developing, modeling will likely focus on the creation of new models, new model components, and/or the development of new codes. Both types of problems are evaluated through model studies in the Delta and are described further below.

Figure 4 diagrams the sequence of steps that might occur for a problem with well-defined basic theoretical principles, mathematical representations, and computer implementations in place. The main steps, explained in greater detail in the following chapters, involve using observed data from the field to configure and calibrate the model; apply to various scenarios; and report results. Model results are compared against field data and can be subjected to a variety of tests to evaluate performance. To provide additional specificity for these modeling best practices, we separate the evaluation step into two phases: an initial evaluation that is expected to be applied for all model applications and additional evaluation such as sensitivity and uncertainty analysis. The latter phase requires more resources and time that are better suited for larger and more consequential exercises. Many applications fall into the category of applications shown in Figure 4, where a modeling framework (such as MODFLOW or C2VSIM, for groundwater flow modeling[1]) is customized for a specific geography. Although the basic theory for this class of models is well-established, there are nonetheless many areas that are the focus of improved performance and research. These include collection of more spatially and temporally resolved field data to better configure the model; improving the calibration of the model to better fit observations; more efficient model run times; improved visualization and interpretation of results; and more sophisticated evaluation of performance as described in the following chapters. Over time, models in this category, while using the same theoretical equations to represent the underlying processes, are becoming more spatially and temporally detailed and resulting in greater computational requirements.

Figure 5 diagrams the sequence of steps that may occur for a problem where the underlying scientific understanding is evolving. The primary difference between an evolving problem and a well-defined problem is that, at the inception of such a study, model structure, data needs, or even outputs are less certain. Here the focus is on collecting more data (typically new types of indicators to improve scientific understanding) and developing conceptual models to explain relevant processes and drivers for a variable of interest. A conceptual model may be thought of as a compact graphical representation of the key processes of interest in a modeling study. Benefits of a conceptual model are described further in Chapter 2. A conceptual model may be converted to a quantitative model structure, thereby formally describing how inputs and outputs are related and then implemented in computer code. Such models may then be calibrated and evaluated in a manner consistent with more mature models. The modeling best practices proposed in this work apply to both newly-defined and well-established modeling processes. The distinction between Figure 4 and Figure 5 is made not to downplay the role of evaluating and testing practices in models with evolving science, but rather to point out that the primary attention may often be focused on improving the basic understanding and representation of the processes of interest.

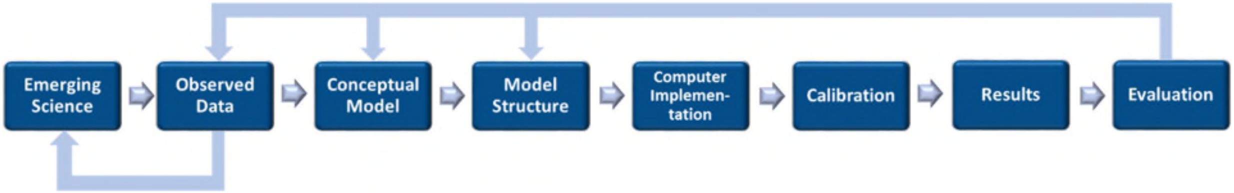
[1] See Memo 1, Model Inventory, for additional details on these and related frameworks.



Areas of Focus:

- More resolved observed data
- Improved model configuration
- Improved calibration
- More efficient performance
- Visualization of results
- Additional evaluation

Figure 4. Modeling steps for a topic with well-developed theoretical frameworks and computer implementation.



Areas of Focus:

- Additional data collection
- Conceptual model
- Model structure

Figure 5. Modeling steps for a topic where the scientific understanding is still evolving.

1.6 Model Life-Cycle

Many models will not be applied to a single study but will have an extended life, either as-is or with modifications and updates, potentially over decades. Alternatively, codes and formulations from one model can be re-purposed and used in a new generation of models. Another perspective to think about modeling best practices, therefore, is over a long-term life cycle. This is shown schematically in Figure 6. The computer implementation of each model is based on a specified conceptual model and model structure. As information from multiple studies applying the model is accumulated, a more nuanced understanding of the strengths and weaknesses of the underlying model structure will develop. This may help to inform and improve the underlying conceptual model, and ideally, result in updates and revisions to the model for future applications. Over the long-term, individuals responsible for model development will likely change and, therefore, there is a need to adequately document the existing model and to develop an effective long-term plan for code maintenance and migration to modern software platforms. The long-term life cycle of the model refers to the activities related to the management and maintenance of the model that enable its continued improvement and evolution over time.

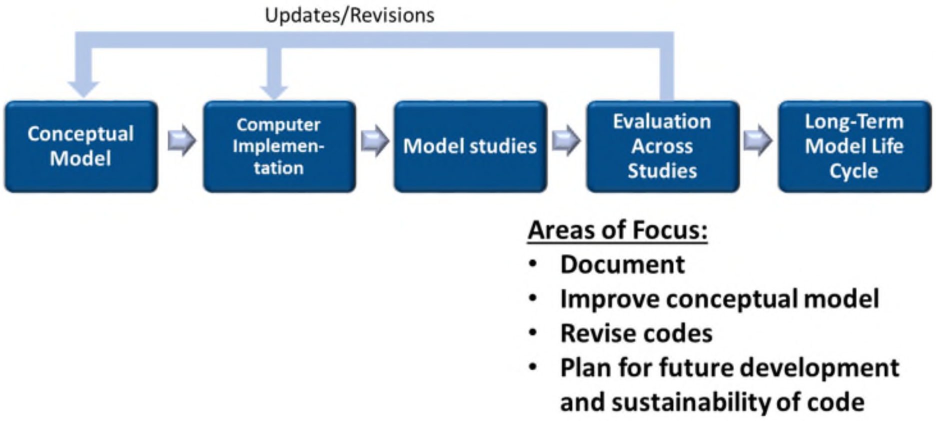


Figure 6. The life-cycle of a typical model.

1.7 Model Ownership

Environmental models commonly used in the Delta may be open source, public domain, or proprietary. Open source models are those where the underlying source code of the model is available for anyone to examine and modify, potentially creating a new executable version of the model. Public-domain models are those where the executable version of a model is freely available, although the source code may not necessarily be available. Finally, proprietary models are owned by a non-public entity and there is a cost for leasing and applying the model. Memo 1 (*Model Inventory*) provides a description of specific models in use in the Delta across different study domains.

Each approach has strengths and weaknesses as outlined below:

- **Open-source models:** These models are free to use and their source codes can be modified by anyone. In many cases, well maintained and documented open-source models may be the basis for major modeling studies, as is the case with the DSM2 and CalSim models in the Delta. Open source models are also suitable for new scientific applications, where there may be a need to add new process information to an existing model by making changes at the computer code level. In most cases, considerable user expertise is needed to make meaningful changes to complex environmental models. Where a community of modeling experts exists, open-source models are an effective means for continued development. In general, however, the ability of any user to change the model can create a concern with version control, in that specific outcomes may be a consequence of the particular variant of the model being used. Furthermore, for open source models to be sustained, there is a requirement for funding of staff for development; often this is done through government or academic organizations.
- **Public-domain models:** These are free to use, although there may be limits to what can be changed in a published form of the model framework. The costs of development are borne by the sponsoring organization. In some cases, sponsoring agencies (e.g., the U.S. Army Corps of Engineers) have provided resources to make their public-domain models easy to use in a manner similar to some proprietary models. These models are suitable for many standardized studies with large teams of modelers.

- **Proprietary models:** Fees for use may be significant, and thus limit who can directly use the model. Fees provide continuing resources for the developing organization to improve the code and the user-friendliness of the model. Where a model has uses in many geographic domains, the development costs are amortized over a larger user base. These models are suitable for standardized studies where available model features adequately represent the modeling purpose, and an off-the-shelf product can be used.

In the Delta, a mix of open source, public domain and proprietary models has evolved in response to several factors, including: the history of development in different domains, sponsoring agency involvement, and resources for new models. As with other elements described in this chapter, the ownership of the model can in some cases influence the best practices actions that can be applied.

1.8 Other Published Modeling Guidance

As the science of modeling has matured and become more widely used, good modeling practices that target particular weaknesses of model development and application have been proposed. Table 2 is a summary of general guidance on environmental modeling that was published over the past two decades and was considered as part of this work. Other domain-specific guidance has been developed and embedded within reviews of modeling studies in different disciplines, e.g., coastal and estuarine models (Ganju et al., 2016; Dawson et al., 2019); watershed models (Daniel et al., 2011); models for total maximum daily load (TMDL) development for water quality constituents (Shoemaker et al., 1997); and modeling nutrient behavior in aquatic systems (Trowbridge et al., 2016). This document is informed by these published guidelines and is tailored to suit the specific modeling needs in the Delta today. Elements from these prior guidelines are cited, as appropriate, throughout the following chapters.

1.9 Motivation for this Work

The goal of this work is to propose a set of best practices for the wide range of modeling activities in the Delta today. Based on the literature cited in Table 2 as well as our own experience, the proposed practices cover both new model development as well as structured applications of existing model frameworks. Furthermore, the proposed practices range from a single study to a model life-cycle over decades. Key benefits of proposing a set of best modeling practices are organized as follows:

Credibility Benchmark: The perceived credibility and resulting acceptance (or rejection) of a model falls between two extremes. At one extreme, model results are accepted indiscriminately without regard to the underlying error or uncertainty in the models. At the other extreme, model results are rejected as being “wrong” because stakeholders do not trust the model or because the model cannot explain all observations. Good modeling practices should assist model users and decision-makers in making informed judgments regarding model credibility.

Table 2. Prior General Guidance for Environmental Modeling

Year of Publication	Title	Author	Focus
2000	*Protocols for Water and Environmental Modeling*	California Water & Environmental Modeling Forum (formerly Bay-Delta Modeling Forum)	Guidance on modeling protocols for the Bay-Delta
2002	*Guidance for Quality* *Assurance Project Plans for Modeling*	U.S. Environmental Protection Agency	Recommendations on how to develop a Quality Assurance Project Plan (QAPP) for projects involving model development or application
2006	*Ten Iterative Steps In Development and Evaluation of Environmental Models*	A. J. Jakeman, R. A. Letcher, and J. P. Norton	Widely cited general guidance on good practices
2007	*Models in Environmental Regulatory Decision Making*	National Academy of Sciences	General guidance on best practices in model use in complex regulatory settings
2008	*Good Modeling Practice*	N. Crout et al.; Chapter in book on Environmental Modeling, Software, and Decision Support, Jakeman et al., Eds.	General guidance on model development, application, and testing
2009	*Guidance on the Development, Evaluation, and Application of Environmental Models*	Gaber et al, 2009 (U.S. Environmental Protection Agency, Council for Regulatory Environmental Modeling	General guidance on environmental models, considering both technical and institutional aspects
2012	*Assessing the Reliability of Complex Models: Mathematical and Statistical Foundations of Verification, Validation, and Uncertainty Quantification*	National Academy of Sciences	Report with a technical focus on analysis approaches for evaluating complex scientific and engineering models

Investment Protection: With the increasing complexity of environmental problems being addressed, model development and related analyses represent a large and growing investment of resources. Unlike databases of field observations, however, model results have limited shelf lives unless supported by adequate documentation, source codes, input files, etc. Good practices should help provide guidance on developing and maintaining such supporting material.

Best Practices Adoption: Our experience tells us that many of the good practices described in the following chapters are often acknowledged by the modeling community but are not fully implemented because of institutional or resource constraints. Our goal in recommending such practices is not to cause an unreasonable burden on model developers, but to highlight a range of realistic options that can be incorporated within ongoing studies. To encourage adoption, we provide a checklist to be used at the inception of a modeling effort and an evaluation form that can be used to evaluate a modeling study following its completion.

Context for Non-Modelers: These best practices are intended to inform the larger modeling community (including model sponsors and stakeholders) so that study results can be reviewed in an inclusive and comprehensive manner. Users of model results, many who may not be model specialists, are confronted with model outputs that may or may not be in an audience-appropriate format. A set of best modeling practices can support basic familiarity (i.e. provide context) with the modeling approaches and limitations, thereby enhancing a users' experience with the model and promote an informed positive vision about the results they observe from the model. Additional investments of resources and time are generally associated with the implementation of best modeling practices; this needs to be communicated adequately to model sponsors and others within the larger modeling community.

1.10 How to use this Memorandum

This document provides guidance that can be used by modelers to support execution of a model study. Many of these concepts are general and can also be considered by non-modelers who need to understand the scope of a modeling exercise at its inception and make a judgement as to the utility of the results upon its completion.

We describe best practice elements under three general themes and devote a chapter to each theme. The first theme, described in Chapter 2, focuses on specific actions to improve the robustness of modeling. These actions apply to virtually all modeling efforts, even where resources are modest and time schedules are limiting. The second theme, described in Chapter 3, focuses on additional steps that are appropriate for large modeling studies with consequential societal impacts, such as planning for major infrastructure or new water quality regulations. In most cases, these steps will require additional resources to perform adequately, and will need to be a topic of discussion between model developers and model users. The third theme, described in Chapter 4, focuses on broader activities associated with modeling that enable better communication and adoption of results. These activities are related less to the technical elements of modeling per se, but to the social elements that ultimately drive a model's utility. These activities should be undertaken over all phases of a model's life, beginning with problem formulation and ending with specific applications and its long-term life-cycle.

To encourage adoption of the best modeling practices described in this document, Chapter 5 provides three documents: i) a checklist to be used at the inception of a modeling study, ii) an evaluation form that can be used to assess and potentially score a model study at completion, and iii) an evaluation form to assess the long-term life cycle of a modeling framework. Terms used in this document are defined in the Glossary.

2 Improve Model Robustness for Typical Applications

Model studies vary greatly in time requirements and resources available for execution. Time requirements may range from weeks to many years, depending on the complexity and the importance of the underlying questions being asked. In this chapter, we identify a set of practices that help to address model robustness and are applicable to virtually all types of modeling activities, including applications that are relatively limited in time and scope. In this work, we refer to robust modeling exercises as those that are credible among the community of modelers and users and stand the test of time.

2.1 Define the Purpose of a Modeling Exercise

At the inception of a study, it is important to clearly define the specific purpose of a modeling exercise. While this practice appears obvious, it is often not explicitly addressed up-front among modelers and stakeholders. A clear specification of the purpose is especially needed for modeling efforts that are not focused on open-ended research. An important goal of this practice is to constrain acceptable outcomes. The stated purpose should, at a minimum, allow stakeholders to agree on a broad scope of work, including: what processes will and will not be modeled, what data are needed, what form the results will take, and what the expected accuracy and uncertainty will be. Importantly, a modeler needs to understand the stakeholders' viewpoint of how the model results will be used. The model study's purpose can be defined with greater clarity when the task at hand consists of customizing an existing framework, rather than creating a completely new model. The more specifics are outlined early, the more efficiently the modeling exercise will progress. In many situations, elements of the modeling scope are not well-defined up-front and are later selected by decision-makers based on the results obtained, likely resulting in a less-than-optimal use of the modeling effort.

A National Research Council (NRC) evaluation on modeling practices for regulatory application (NRC, 2007) proposed the following relevant and valuable suggestions to help define the model purpose. Not all of these questions may apply to all modeling efforts, but a reasonable subset can be selected for most modeling studies:

- At what temporal and spatial scales is the model to be applied? This question involves the grain (resolution in time and space) and the extent (spatial and temporal domain) at which the model is to be focused.
- Who will the major model users be and what constraints does that imply for model application once developed? What is the level of expertise of the proposed users?
- What type of input data must the model users provide? How can these data be obtained (from other models and measurements)?
- What sources of data are available to support model evaluation?
- What are the basic outputs needed and must they be constrained by a deterministic approach or is a probabilistic approach allowable[2]? What additional outputs might be useful to enhance model transparency (e.g., enhance ability to explain findings to stakeholders and users) and flexibility (e.g., capacity for the model to be modified and applied to situations for which it was not constructed)?
- What level of reliability is required?
- What evaluation criteria should be applied to determine the applicability of the model or of particular model components?
- The exercise to define the model's purpose should be formally and clearly documented, a common engineering practice across many other disciplines (e.g. civil, environmental, mechanical) in their design endeavors. That documentation can then be revisited and revised as the modeling effort evolves.

2.2 Develop Conceptual Models and Transparent Model Formulations

Conceptual models, as used in the environmental domain, are abstractions of reality, ranging from a schematic representation of processes to a more detailed description of the state of science related to a specific environmental concern. Where a new model is to be developed, creating a conceptual model, even a simple schematic representation, is recommended as a first step in creating the model and writing documentation. A conceptual model is a communication tool that guides model development, experimentation, and evaluation. Moreover, conceptual models may provide a good tool for communicating about a model with stakeholders, particularly when the conceptual model represents processes graphically and highlights key quantitative information. A good conceptual model improves understanding of the system and creates a point of reference for model developers to revisit when considering changes to the model.

[2] A deterministic output provides a single set of model results, potentially varying in space and time. A probabilistic output provides a distribution in output values based on the input conditions and parameter values used.

In addition to enhancing communication, under certain circumstances, a well-designed conceptual model may more readily accommodate formal hypothesis-testing relative to a computer implementation of a conceptual model. A good conceptual model may lead to the early realization that development of a quantitative system model would be premature due to data and knowledge gaps.

In some instances, conceptual models play a role following the synthesis of data and after completion of a modeling study. Typically, the initial conceptual model would be refined over the course of model application, and more quantitative information provided in the revised conceptual model. Such a model can serve as a basis for further communication with stakeholders (also see Chapter 4). Graphical representation of modeled processes, with key quantitative information being highlighted when available, is a significant aid to communicating with stakeholders.

Three examples of Delta mercury conceptual models highlight the roles played by different types of conceptual models at different stages of planning and model development. The first example, from Wiener et al. (2003), is shown in Figure 7 as a graphical summary of all the mercury transport, transformation, and bioaccumulation processes from the upper watershed to the estuary. This graphic illustrates the complexity of interactions in different geographical areas within the watershed and points to the need for monitoring, analysis, and possible modeling needs for different conditions. It does not provide specific process level information that a modeler could use, however. A more focused conceptual model on freshwater cycling of mercury (Figure 8), from Hudson et al. (1994), provides such process detail and was used in the formulation of a numerical model (the Mercury Cycling Model). A third example, from Wood et al. (2006), provides a conceptual model with quantitative information on the Delta methylmercury budget (Figure 9) that was compiled using available data on concentrations and flows. Extensive new work on mercury modeling in the Delta (see Memo 2, Chapter 12) builds upon these conceptual representations. While the new modeling may result in changes to the load estimates and to the conceptual models, the existing conceptual models remain an important communication tool and the basis for collective understanding among the stakeholder community.

Conceptual models have been developed and documented as stand-alone volumes, combining graphical representations and narrative syntheses of available information. Examples of conceptual model documentation include a model for Delta Smelt (Interagency Ecological Program, 2015) and a model for nutrients in the Central Valley and Delta (Tetra Tech, 2006). These detailed conceptual models are applicable when a large amount of information exists on a problem of interest, and where available, provide a strong foundation for model development.

A potential pitfall associated with conceptual models is that the model may be overly abstract to sufficiently guide the implementation of a quantitative system model. This problem is particularly relevant for large, complex processes. As a countermeasure, conceptual models should be re-evaluated and revised as the quantitative model is developed.

2. Improve Model Robustness for Typical Applications

Another potential pitfall associated with conceptual models is that the model may be too complicated to allow prioritization of key processes, is overly difficult to understand, and overly difficult to communicate to the modeling team and to stakeholders. In developing conceptual models, i) professional judgement is needed to identify and prioritize key processes and mechanisms and ii) assumptions must be clearly documented.

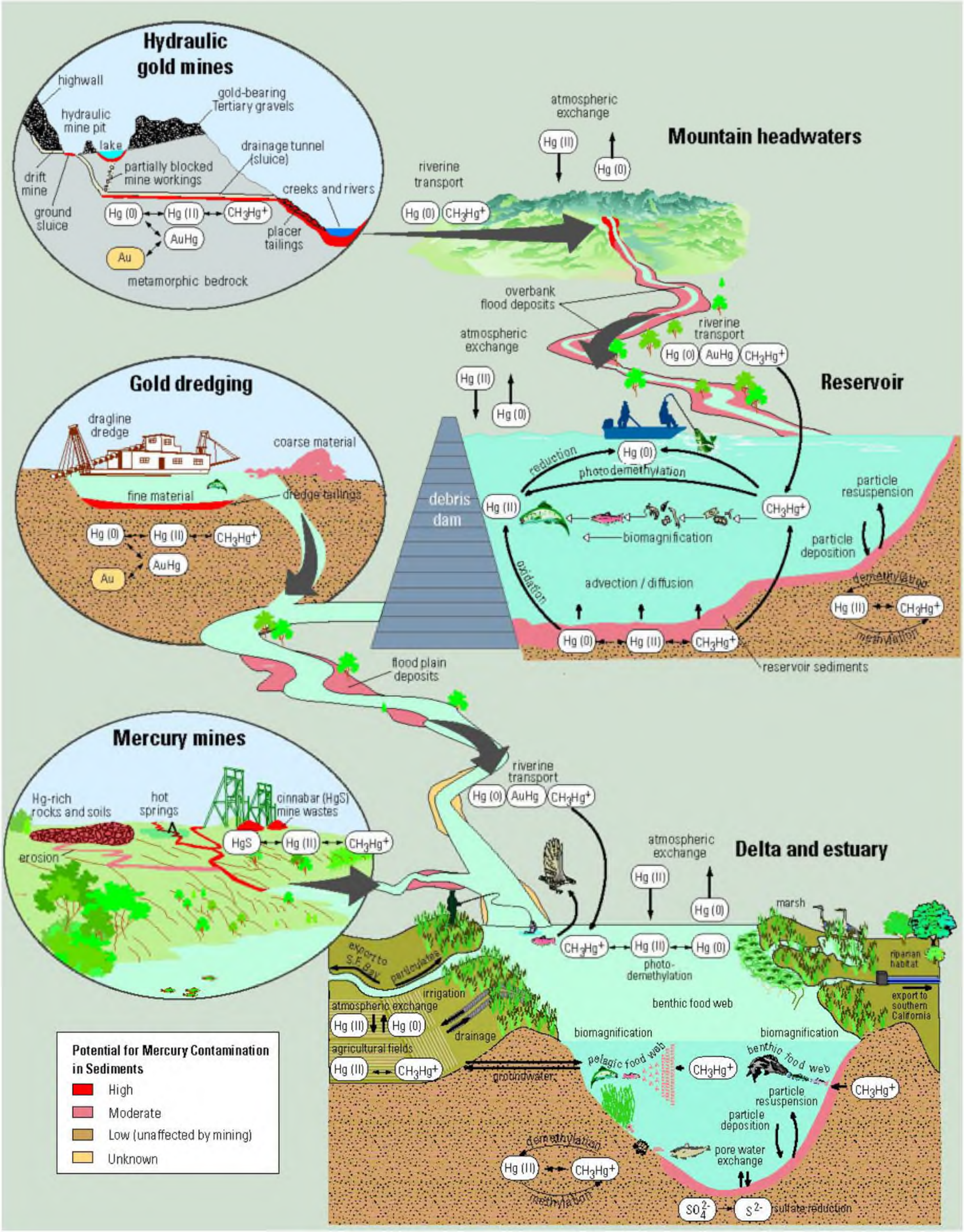


Figure 7. Conceptual model for mercury transport and biogeochemistry in Bay-Delta ecosystem. Source: Wiener et al. (2003).

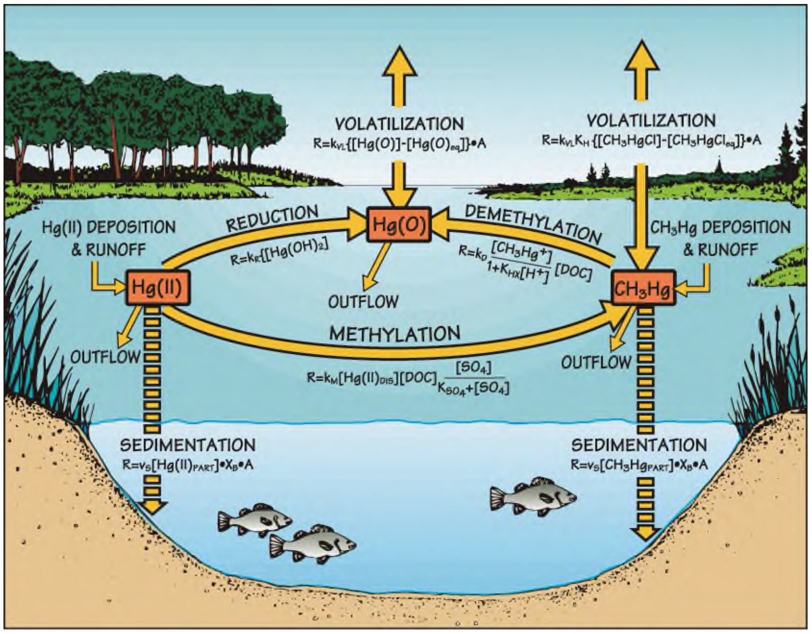

VOLATILIZATION
$R=k_{VL}\{[Hg(O)]-[Hg(O)_{eq}]\}\bullet A$
VOLATILIZATION
$R=k_{VL}K_H\{[CH_3HgCl]-[CH_3HgCl_{eq}]\}\bullet A$
Hg(II) DEPOSITION & RUNOFF
REDUCTION
$R=k_R[[Hg(OH)_2]$
Hg(O)
DEMETHYLATION
$R=k_D\frac{[CH_3Hg^+]}{1+K_{HX}[H^+]}[DOC]$
CH₃Hg DEPOSITION & RUNOFF
Hg(II)
OUTFLOW
CH₃Hg
METHYLATION
OUTFLOW
OUTFLOW
$R=k_M[Hg(II)_{DIS}][DOC]\frac{[SO_4]}{K_{SO_4}+[SO_4]}$
SEDIMENTATION
$R=v_S[Hg(II)_{PART}]\bullet X_B\bullet A$
SEDIMENTATION
$R=v_S[CH_3Hg_{PART}]\bullet X_B\bullet A$

Figure 8. Conceptual model for mercury cycling reactions in freshwater systems (Hudson et al., 1994).

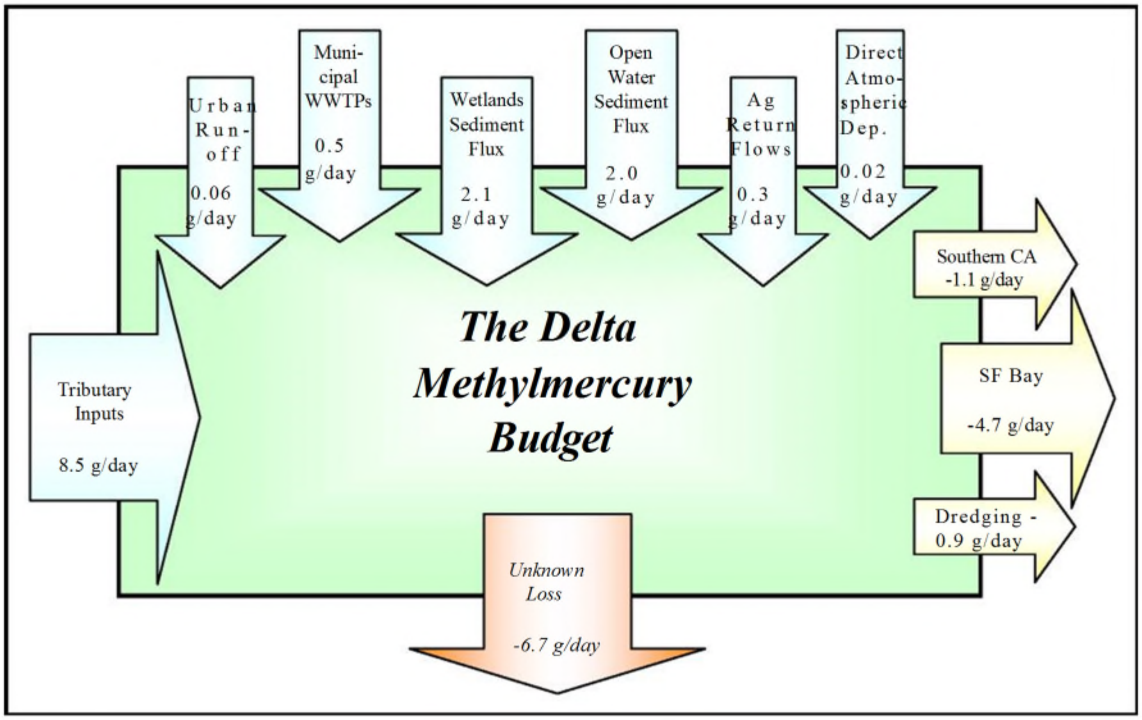

Urban Run-off 0.06 g/day
Municipal WWTPs 0.5 g/day
Wetlands Sediment Flux 2.1 g/day
Open Water Sediment Flux 2.0 g/day
Ag Return Flows 0.3 g/day
Direct Atmospheric Dep. 0.02 g/day
Southern CA -1.1 g/day
The Delta Methylmercury Budget
Tributary Inputs 8.5 g/day
SF Bay -4.7 g/day
Dredging - 0.9 g/day
Unknown Loss -6.7 g/day

Figure 9. Methylmercury daily budget based on load analyses in the Delta methylmercury TMDL. Source: Wood et al. (2006).

2.3 Verify Code

Code verification is the process of determining how accurately a computer program correctly solves the equations of a mathematical model. It is assumed that most established model frameworks in common use will have undergone this test and, thus, this task is appropriate when a new code or module is being developed for a specific application. Code verification also provides an opportunity to evaluate or reevaluate the efficiency of the code, which may enable its use for situations that require multiple model runs, such as for sensitivity analysis. Typically, computer codes are verified with well-documented data sets and the results of published and documented analytical or semi-analytical models. Within many large-scale computational models, opportunities exist to perform verification studies that reflect the hierarchy or collection of these models. For example, code verification can successfully employ "unit tests" that assess whether the fundamental software building blocks of a given code correctly execute their intended algorithms. Documentation of code verification, especially for newer models or for models where modifications are being made to established codes, is an important part of establishing model robustness.

2.4 Prepare Standardized Observed Datasets for Analysis

Observed data are a fundamental part of sound modeling practice. In most instances, environmental models contain parameters that are defined independently or are adjusted as part of the model setup (see Chapter 1). Parameters that cannot be determined independent of the model (e.g. a reaction rate for a chemical process within a water body or a roughness coefficient for a stream bed) are estimated through the process of model calibration (described below), which involves tuning the parameters to obtain a good fit between the model and observed data. Thus, data that are credibly measured, have good quality, have been cleaned of potential erroneous values, and well documented for limitations are an essential part of the modeling process. California Assembly Bill 1755 (Open and Transparent Water Data Act, AB 1755) is a large step toward such a data resource. The bill requires California state agencies to make data publicly available and to develop protocols for data sharing, documentation, quality control, and promotion of open-source platforms and decision support tools related to water data. Once fully implemented, AB 1755 may provide observed data in a form that is suitable for model studies (i.e., for calibration and testing). Additionally, the California Water Quality Monitoring Council has requirements for quality assurance program plans that must be used in collecting water quality data.

Modelers generally prepare data sets for model calibration and testing, pulling from a variety of available data sources. Data preparation typically involves some form of compilation across different sources, conversion to common units, and cleaning to remove known outliers, all of which can be time consuming, and more importantly, have a bearing on the model calibration. Modelers may obtain different model calibration result depending on the quantity of data used and the specific process steps used to prepare model input data, even when utilizing the same model framework. In some cases, models may use processed values (e.g. loads derived from pollutant concentrations, or salinity isohalines from point-based values) rather than directly observed data. The creation of standardized input datasets — whether using directly observed data or some processed form — is recommended, especially when many different users are expected to be involved in parallel studies.

Standardized datasets should be thoroughly described via metadata, corroborated with explicit references, and prepared for analysis with a reproducible and documented workflow. The use of standardized datasets limits the time lost because of errors in model runs arising from incorrect input data. Preparation of standardized datasets involves decisions about missing values and specifications of data types (e.g., date, integer, string, etc.). One of the central challenges of preparing standardized datasets is anticipating the possible ways that a dataset could contain inconsistencies. For example, does the data provider use letter codes in place of missing data causing potential type mismatches? Does the test dataset include all the possible permutations of codes produced by a data provider? Does the data cleaning code check for the shape of the data for early detection of possible changes in data structure? Ideally, these decisions should be documented for future users.

Typically, such processed datasets are not easily available and may not be part of raw data sources. Therefore, a related recommendation is the creation of a searchable repository of such standardized datasets where a user can identify appropriate information for use in a modeling study.

2.5 Accommodate Appropriate Model Complexity

In most modeling situations encountered in the Delta, more than one approach may be taken for model development. Where the task requires use of an existing model, more than one model may be available for use. Where the task requires the development of a new model, the modeler has some discretion on the level of process complexity to be used. In both cases (existing or new model), the modeler has the flexibility to determine the level of spatial and temporal detail incorporated. As an example, a dynamic model may compute and report values at timesteps of minutes, days, or longer. A spatially detailed model may contain a grid with sizes ranging from square meters to hundreds of square kilometers. An appropriate level of model complexity is a function of multiple factors including, but not limited to, objectives of the modeling exercise, knowledge of the system, and data availability. A useful rule of thumb for deciding on the level of complexity *a priori* is that the model outcomes should be testable by observed data spatially and temporally. For example, when choosing between a simple lumped model and a more sophisticated distributed model, a model developer should be able to 1) support the added complexity by more detailed input data available for the distributed model (e.g. distributed measurement of related properties), and 2) test whether this added complexity is providing an additional benefit by comparing the simulations with available observed data.

Inability to identify a single correct model (i.e. problem of identifiability) is a clear cost of excessive model complexity. Complex models often have a larger number of parameters, and under these conditions, different combinations of parameter values can lead to similar model results when compared to observed data (e.g., runoff at the catchment outlet or water level in groundwater bores). Such a result implies that the observations are insufficient to properly test the model structure or parameter values. Furthermore, even if a model appears to accurately simulates a particular response, this result does not necessarily indicate that other model predictions are correct. For example, although a rainfall-runoff model may provide good fits to streamflow at a catchment outlet, it may not necessarily provide accurate streamflow estimates at internal gauging stations or correct spatial patterns of saturation deficit. This issue has been clearly identified by many researchers (Grayson and Blöschl, 2001; Tasdighi et al., 2018), yet it is commonly ignored by model users. This issue is often referred to as "equifinality" or "non-uniqueness" in the literature and is a subject of continuing discussion (Beven, 2001).

Figure 10 illustrates the conceptual relationship between model complexity, data availability, and predictive performance. The term "data availability" refers to both the amount and quality of the data in terms of its use for model testing. Within the context of hydrology, access to spatial patterns of surface runoff data is considered "high" availability while scarce streamflow measurements as aggregated runoff implies "low" availability. The term "model complexity" means detail of process representation and spatial/temporal detail. Complex models include more processes and report values at greater spatial and temporal density. As illustrated in Figure 10, for a given data availability, there is an optimum level of model complexity giving the highest predictive performance; additional complexity leads to concerns with identifiability or equifinality. For a given model complexity, more data availability usually results in better predictive performance up to a point, beyond which the data does not provide more useful information to improve the model with that level of complexity. Under these conditions, a model user may wish to consider a more complex model to better exploit the information from the available data.

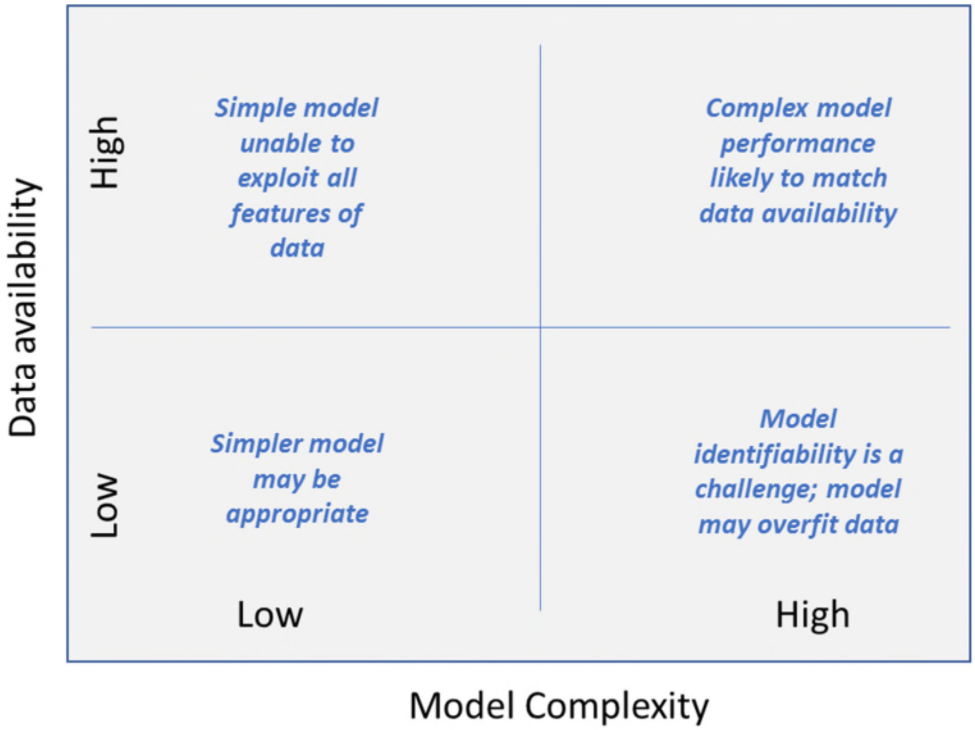

Data availability
High
Low
Simple model unable to exploit all features of data
Complex model performance likely to match data availability
Simpler model may be appropriate
Model identifiability is a challenge; model may overfit data
Low
High
Model Complexity

Figure 10. The conceptual relationship between model complexity, data availability, and performance (modified from concepts in Grayson and Blöschl, 2001)

2.6 Calibration: Formal Process for Parameter Estimation in Models

As previously discussed, environmental models often use parameters that are not known ahead of time but are derived on a site-specific basis from the observed data. Models use parameters within equations to relate various influences and responses (e.g., rainfall to runoff). Some of these parameters may be readily determined based on field measurements or other observations. Often, however, many model parameters are either too difficult to measure (specifically with proper spatial resolution) or practically impossible to measure (non-measurable parameters). An example of a parameter that is too difficult to measure with adequate spatial resolution includes the hydraulic conductivity in aquifers (used for groundwater modeling); or the parameter Manning's n coefficient for roughness in surface water bodies (used for streamflow modeling). Furthermore, some domains, notably in the biological, economic, and social sciences, inherently use parameters that are lumped and location specific, and not known *a priori*.

Depending on the level of complexity, models can be posed with a small number of parameters or can be posed with a very large number of parameters – in extreme cases numbering in the thousands. The task of calibration—also termed training—is to find the set of best-fit parameters that describe the observed data with a given model. Formally, calibration is the mathematical process of searching for a solution that minimizes or maximizes an objective function (i.e. a function quantifying a measure of error based on model simulations and observed data), by adjusting the values of *n* unknown parameters, i.e., a search in *n*-dimensional space. The general goal is to find a global best-fit, but in complex models this is often difficult, and it is not uncommon to find model calibration codes settling in local minima. Superficially, local minima have some features of a global minimum, but formally, they do not represent the best parameter fit.

There is a wide range of objective functions commonly used in the literature for model calibration and testing. Selecting an appropriate objective function for model calibration and testing has been a subject of continuing discussions in the literature of environmental modeling. Table 3 presents a list of common model performance metrics. Since all model performance metrics have strengths and weaknesses, it is recommended that more than one metric (i.e., multi-objective optimization) be considered for calibration/testing of models. However, care should be taken as these metrics have different units and ranges. There are numerous published algorithms to help perform this search that are used in conjunction with environmental models, of which the Parameter Estimation and Uncertainty Analysis (PEST) tool is widely used for environmental models (theory in Doherty and Hunt, 2010; Doherty, 2015; example application in Doherty and Johnston, 2003). A list of widely used model sensitivity, calibration, and uncertainty analysis frameworks along with brief description is presented in Chapter 3.4.

The search process of finding best-fit parameters in calibration requires the model to be run multiple times, each run using a new combination of parameter values. As the number of parameters in a model grows, and as the model run-time increases, the computational burden of automated calibration grows exponentially. In many cases where complex, computationally intensive models are being used (with single run times over hours to days), calibration is often a more manual process, with expert users interacting with the model and applying knowledge of the parameter space to tune the overall performance. In a manual calibration process, model parameters are essentially tuned to minimize the difference between the model simulation and observed data. This is an iterative procedure and usually several rounds of model runs are performed to locate parameters that mimic the observed data with reasonable accuracy. Alternatively, additional computer resources are deployed during the calibration period, running the model on supercomputers or on the cloud to circumvent the computational burden.

Table 3. Common Model Performance Evaluation Metrics

General category	Performance metric	Description	Issues	Reference
Standard Regression	*Slope and y-intercept*	The slope indicates the relative relationship between simulated and measured values. The y-intercept indicates the presence of a lag between simulated and measured data, or that the data sets are not perfectly aligned. A slope of 1 and y-intercept of 0 indicate that the model perfectly reproduces the measured data.	Most often the underlying assumptions of linear regression (normality, randomness, etc.) are overlooked which can undermine the credibility of the inference from a regression model	Willmott, 1981

General category	Performance metric	Description	Issues	Reference
	Pearson's correlation coefficient (r) and coefficient of determination (R2)	r and R2 indicate the degree of collinearity between simulated and measured data. r, is an index of the degree of linear relationship between observed and simulated data and ranges from −1 to 1. If r = 0, no linear relationship exists. If r = 1 or −1, a perfect positive or negative linear relationship exists. Similarly, R^2 describes the proportion of the variance in measured data explained by the model. R^2 ranges from 0 to 1, with higher values indicating less error variance, and typically values greater than 0.5 are considered acceptable.	r and R^2 are very sensitive to high extreme values (outliers) and insensitive to additive and proportional differences between model predictions and measured data.	Santhi et al., 2001
Dimensionless	*Index of agreement (d)*	Standardized measure of the degree of model prediction error and varies between 0 and 1. A computed value of 1 indicates a perfect agreement between the simulated and measured values, and 0 indicates no agreement at all.	d is overly sensitive to extreme values due to the squared differences.	Willmott, 1981
	Nash-Sutcliffe efficiency (NSE)	The Nash-Sutcliffe efficiency (NSE) is a normalized statistic that determines the relative magnitude of the residual variance (“noise”) compared to the measured data variance (“information”). NSE ranges between −∞ and 1.0 (1 inclusive), with NSE = 1 being the optimal value. Values between 0 and 1.0 are generally viewed as acceptable levels of performance. Values <0 indicate that the mean observed value is a better predictor than the simulated value, indicating unacceptable performance.	NSE is sensitive to high extreme values.	Nash and Sutcliffe, 1970
	Persistence model efficiency (PME)	PME is a normalized model evaluation statistic that quantifies the relative magnitude of the residual variance (“noise”) to the variance of the errors obtained by the use of a simple persistence model. PME ranges from 0 to 1, with PME = 1 being the optimal value. PME values should be larger than 0.0 to indicate “minimally acceptable” model performance.	Explicit assumption that variance increases linearly with time which should be revisited depending on the problem	Gupta et al., 1999

General category	Performance metric	Description	Issues	Reference
	Prediction efficiency (Pe)	P_e is the coefficient of determination (R^2) calculated by regressing the rank (descending) of observed versus simulated constituent values for a given time step. P_e determines how well the probability distributions of simulated and observed data fit each other. A prediction efficiency of 1 is perfect agreement at all times. Prediction efficiencies less than or equal to 0 do not provide useful predictions of the time variation of the observations.	Sensitive to high extreme values	Santhi et al., 2001
	Performance virtue statistic (PVk)	The performance virtue statistic (PV_k) is the weighted average of the Nash-Sutcliffe coefficients, deviations of volume, and error functions across all flow gauging stations within the watershed of interest. PV_k can range from $-\infty$ to 1.0, with a PV_k value of 1.0 indicating that the model exactly simulates all three aspects of observed flow for all gauging stations within the watershed.	Since the main criteria used is NSE, this metric can also be prone to biases from large error residuals	Wang and Melesse, 2005
	Logarithmic transformation variable (e)	The logarithmic transformation variable (e) is the logarithm of the predicted/observed data ratio. The value of e is centered on zero, symmetrical in under- or overprediction, and approximately normally distributed.	Not widely used and may not add much value considering the underlying distribution	Parker et al., 2006
Error Index	*Mean absolute error (MAE), Mean square error (MSE), and Root mean square error (RMSE)*	RMSE, MAE, and MSE values of 0 indicate a perfect fit. RMSE and MAE values less than half the standard deviation of the measured data may be considered low and that either is appropriate for model evaluation.	Since these metrics use averaging on error residuals, they may not be suitable as an objective function for calibration. However, they can be used as additional performance validity metrics once the model is calibrated.	Moriasi et al., 2007

General category	Performance metric	Description	Issues	Reference
	Percent Bias (PBIAS)	Percent bias (PBIAS) measures the average tendency of the simulated data to be larger or smaller than their observed corresponding values. The optimal value of PBIAS is 0.0, with low-magnitude values indicating accurate model simulation. Positive values indicate model underestimation bias, and negative values indicate model overestimation bias.	The effects of individual error residuals may smooth out due to averaging	Gupta et al., 1999
	RMSE-observations standard deviation ratio (RSR)	RSR standardizes RMSE using the observations standard deviation, and it combines both an error index and the additional information. RSR is calculated as the ratio of the RMSE and standard deviation of measured data. RSR varies from the optimal value of 0, which indicates zero RMSE or residual variation and therefore perfect model simulation, to a large positive value. The lower RSR, the lower the RMSE, and the better the model simulation performance.	Same issues with RMSE	Gupta et al., 1999

Another fundamental challenge associated with model calibration is that the relationship between model error and fitting parameters (termed the error surface) may be complex and fitting procedures may produce locally-optimum rather than globally-optimum parameter values. Local and global minima for a single variable are shown conceptually in Figure 11. Parameter identifiability is the possibility of learning the true values of underlying parameters with a large experimental dataset (Raue et al., 2009). Parameter identification for complex models is very challenging and true parameters values are often not obtained because of the increased computation burden. The topic of parameter estimation in environmental models is an active area of research, focusing on improving efficiency in search strategies and on finding global best fits (Solomatine et al., 1999; Thiemann et al., 2001; Madsen, 2003; Zhang et al., 2011; van Vliet et al., 2016). Regardless of the approach used for calibration, model documentation should describe the approach and explain why the approach is credible for a specific model.

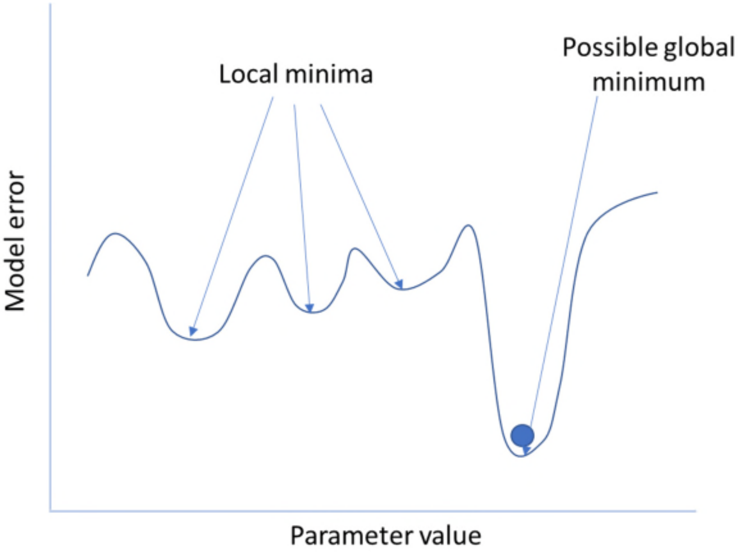


Figure 11. Schematic representation of a complex error surface with multiple local minima.

2.7 Quantitative Model Evaluation and Validation

The term validation has traditionally referred to the process of comparing model predictions with a data set that is independent of model calibration. While this process can provide a generally reasonable evaluation or assessment of the model performance, interpretation of validation results should be developed with care. All models contain inherent uncertainty, which makes the term "validation" somewhat of a misnomer. Because of the term's root (i.e. "valid"), the process of model validation implies an unjustified assertion of legitimacy (e.g. Oreskes and Belitz, 2001). According to this argument, the use of the word validation is misleading in the context of assertions or implications that models accurately reflect underlying natural processes and can be used to provide reliable input for policy and decision making. For example, there is unavoidable uncertainty associated with the subjectivity of what constitutes acceptable error (Konikow and Bredehoeft, 1992).

More generally, the term model validation, as used here, refers to an idea or hypothesis of how a system works or operates, expressed quantitatively. When considering the process of modeling and its evaluation, assessment and value, it is important to note that any physical theory is always provisional. "No matter how many times the results of experiments agree with some theory, you can never be sure that the next time the result will not contradict the theory. On the other hand, you can disprove a theory by finding even a single observation that disagrees with the predictions of the theory (Hawking, 1988). Site-specific hydrologic and ecosystem models are elements of applied science -- in effect, an agglomeration of multiple physical, chemical and biological theories. As such, they are subject to improvement via invalidation, but cannot be proven valid because validation does not necessarily add to the fund of knowledge.

The above theoretical and philosophical discussion is important and environmental modelers should be aware of the fundamental limitations of the term "model validation." A more appropriate term that may be used for describing model performance with respect to observational data is “evaluation” or “testing”, considering not only the modeling fit but also the underlying assumptions that led to that outcome (NRC, 2007). However, we recognize that in recent years, the term model validation has come to be defined more narrowly in the environmental, engineering and modeling communities: as the assessment of a model’s predicted performance against a set of field data, where the model has been calibrated using an independent set of field data. When extended to the non-modeling community and stakeholder, a more qualified approach would help in presenting the model outcomes in the context of model limitations, assumptions and uncertainty.

Given the above broader considerations on terminology, at a more practical level, a common framework may be used for evaluating model results in a systematic manner. A range of visualization approaches (one or more of the combinations shown in Figure 12) is considered suitable for evaluating quantitative results of the performance of a previously calibrated model. A model’s target performance may be defined as part of the stated modeling purpose or based on the best professional judgment of the modelers, given the uncertainties in input data, model parameters, and model structure.

2.8 Non-Quantitative Model Evaluation and Validation

Going beyond the evaluation of numeric results and data as identified in the previous section, given our experience with environmental models, a broader set of considerations may be applied:

- **The bias of stasis.** When model parameters are adjusted to obtain a best fit with historical data, a bias is created towards existing trends, even when driving forces indicate that the model will diverge from existing conditions. A relevant example for modeling hydrology in California and the Delta is hydrologic effects from climate change. Changing climate is affecting driving forces including evapotranspiration trends, the frequency and magnitude of precipitation events, and changes in groundwater and surface water pumping in response. These driving forces are outside the typical range in the historical record and thus the future conditions may diverge outside of a model built on past conditions. It will often be important to run models with varying assumptions about driving forces.
- **Capturing causal processes.** If the underlying causal processes are important, as is the case with most hydrologic and ecosystem models, the model must capture them to be reliable. Model post-audits (see next chapter) can serve to provoke curiosity about why the model does or does not make accurate predictions.
- **Conceptualization.** Models may match observations but still be conceptually flawed. Advances in computational power may help with this in the sense that if a model is run using an exhaustive sampling of parameter values and comes up short, a conceptual error is likely. This is perhaps the thorniest of modeling issues and a difficult one to address, and the reason post-audits are so important.

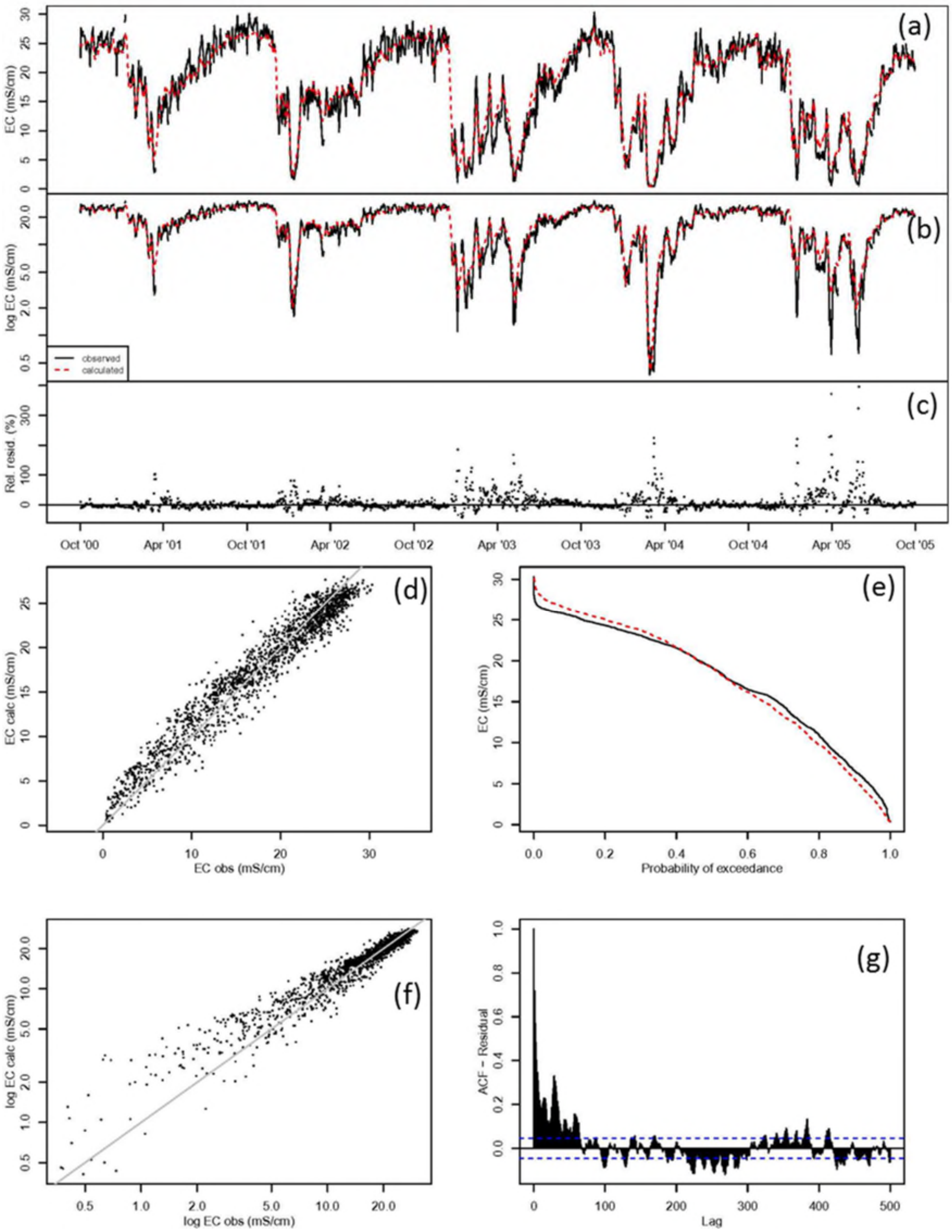


Figure 12. Visualization of adequacy of model performance. Following Crout et al. (2008), but applied to salinity at Martinez in the western Delta, using observed data (Hutton et al., 2015) and a published model of salinity (Rath et al., 2017). (a) Linear time series plots of data and observations (solid line: model; dashed line: observations), (b) log-scale time series plots, (c) plot of residuals (difference between modeled and observed values), (d) observed versus modeled data on a linear scale, (e) cumulative distribution function of observed and modeled values, (f) observed versus modeled data on a log scale, and (g) autocorrelation function of residuals.

- **Overparameterization.** The level of model complexity and number of parameters should be commensurate with the available data and required predictive resolution. Model developers refer to the concept of "model parsimony," which is the development of models with the least number of parameters that adequately explain a relevant phenomenon. While more complex models can be made to fit observed data, this "over-fitting" approach may result in limited model ability to generalize.

Model "validation" should not only include a post-audit but also a close look at the model conceptualization, the ability to capture causal processes, biases (including biases towards stasis), and parameterization.

2.9 Model Documentation for Users and Developers

Documentation may apply to both a general modeling framework and to a specific application. With respect to a framework, in many cases model frameworks are developed and maintained over years, sometimes with different individuals or teams with changing member composition. Good model documentation should serve the needs of developers and users and may be accomplished by using the same set of documents. Ideally, documentation should be prepared in a manner that contains enough information to allow for the long-term evolution of a model, both within the organization and external to it. From the perspective of external users in particular, documentation should explain the basis of the model and its use, including how key input variables are selected. Such documentation should include representative input files and result files to allow a user to reproduce a basic set of scenarios.

In the case of a specific application, documentation needs to be oriented toward explaining the best practice elements that are outlined in this section, including, model purpose, input data used, calibration approach, model evaluation, and model results in the context of the intended purpose.

Writing model documentation is an essential step in model development. However, under short timelines and tight budgets, preparing documentation may become a low priority, particularly for models developed for a specific application with no expectation of re-use. Missing, inadequate, or out-of-date documentation is a barrier to model integration and may result in duplication of effort because a potentially suitable model may be overlooked for inclusion in an integrated modeling process.

Documentation can be broadly classified as internal and external. Internal documentation is generally embedded within the code in the form of function descriptions, code comments, etc. Internal documentation is important and should follow the conventions of the programming language(s) used to build to the model. External documentation is generally written for three audiences: (1) the developer(s) building, maintaining, and updating the model, (2) other modelers interested in the details of the model, including those interested in integration, and (3) users of the model with no need to understand the inner details of the model.

Writing documentation that is easily understood across modeling domains is one of the challenges for model integration. The following elements are recommended to produce good documentation that facilitates integration:

- A general description of the model that includes the modeling goals and the scope of the model.
- A point of contact for the model and information about how to get started using the model, including download links, installation instructions, hardware requirements, and licensing costs (if applicable).
- Assumptions and limitations of the model.
- Model relationships and mathematical methods used.
- Data used to inform model relationships and input data requirements to run the model, including example input files.
- Model output format(s), including example output files.
- Representation of uncertainty.
- Availability of tools for conducting sensitivity analysis, post-processing results, etc.
- Table(s) with all model parameters and their default values.

A recent set of documents prepared for the California Central Valley Simulation Model (C2VSIM; provided online at https://water.ca.gov/Library/Modeling-and-Analysis/Central-Valley-models-and-tools/C2VSim) is an excellent example of documentation addressing most of the questions above, and serving a range of audiences.

2.10 Summary

This chapter provides a brief summary of actions that need to be undertaken to improve the robustness of virtually all modeling exercises, beginning with the definition of the modeling purpose, developing conceptual models to communicate with stakeholders and provide an understandable version of the model, preparation of standardized datasets that can be used to replicate a modeling study and compare across models, verification of code to ensure that the theoretical framework has been correctly implemented (especially true of newly developed models), an adequately documented calibration process to obtain adjustable parameters, and effective visualization of model performance over new data sets. The term validation is often used to describe the testing of model performance with new data, although a more neutral term, evaluation, may be preferred because the typical process of model testing is always bound to be limited to a range of data, and few models can be truly considered valid under all conditions. If a modeler chooses to use the term validation — consistent with the science and engineering literature—it is important to communicate to non-specialists its more narrow definition associated with modeling practice.

In addition, these best practices also suggest a broader exploration of model structure and bias, going beyond the routine calibration and evaluation/validations exercises, especially when observed data do not match model predictions. Finally, for the long-term utility of a model, it is essential that adequate documentation be developed, meeting the current and future needs of users and modelers. Additional steps can be taken to improve the robustness of modeling exercises, but it may not be practical to require them for all studies; a list of such practices for more complex models is described in the following chapter.

3 Improve Model Robustness for Key Applications

The previous chapter summarized actions that need to be undertaken to improve the robustness of virtually all modeling exercises. Additional actions are recommended beyond those previously summarized when model applications entail greater complexity (e.g. integrated models) and/or when model applications are used to support critical decision making that will have significant societal consequences in terms of benefits, costs, and risks. These additional actions will almost always require more time and resources to complete, and this should be clearly scoped with the model sponsors and stakeholders. While greater costs are associated with these actions, the resulting model outcomes will be more robust and more generally accepted by the modeling community and model users.

3.1 Peer Review

Peer-review is the process of soliciting input from experts who are not involved in a particular study but are familiar with the general topic. Peer review should provide timely, open, fair and helpful input and should ideally occur at various stages of the modeling life cycle, including conceptual model development, model implementation in code, and model application to specific geographic area or problem.

Experience indicates that peer review is most helpful when the following conditions are met: i) peer review is conducted in an atmosphere of transparency, collaboration and shared sense of purpose; ii) the review team reviews the source material and modelers' responses to their comments; iii) adequate time and funding is budgeted for review; and

iv) the review team contains some interdisciplinary membership to allow for a broader evaluation of basic assumptions and utility of the exercise. If a sincere commitment to obtaining constructive feedback is not made through the above steps, there is a risk that peer-review becomes more of a rubber-stamp than a positive contribution to a modeling study. Normally, a peer review can increase acceptance of a project.

We recommend that most complex and consequential model studies in the Delta be subject to peer review. Usually such reviews are conducted by the organization sponsoring the model study. For newly developed model frameworks, the process of anonymous peer review required by scientific journals serves as the touchstone for validation of a modeling study and is also recommended.

3.2 Sensitivity Analysis

Sensitivity analysis explores how changes in model inputs—most generally, boundary conditions, parameters, or configuration (as shown in Figure 3)—affect the variation in model outputs. Sensitivity analysis can illustrate which parameters have the least effect on results of interest, and in some cases, may allow for reduction of model complexity, by streamlining process representation. A related concept is uncertainty analysis, where model inputs are presented in a probabilistic form (i.e., as a distribution of values based on current information) to a calibrated model and the effects on model output are evaluated as shown in Figure 13. Sensitivity analysis also complements model calibration, which involves selecting parameter values based on the fit between model output and actual observations. Performing sensitivity analysis after model calibration helps to identify which fitted parameters are close to optimal estimate because low sensitivity indicates high uncertainty in the fitted parameter estimate. Both sensitivity and uncertainty analysis require the running of a model multiple times with a range of inputs. Specific steps for uncertainty analysis are described in the following section.

Sensitivity analysis is often used prior to conducting uncertainty analysis to increase the efficiency of uncertainty analysis by reducing the dimensionality of the model. Using sensitivity analysis, the modeler determines which model parameters have the highest impact on simulations (Saltelli et al., 2008). This will help the modeler to decide which model parameters should be included in the uncertainty analysis procedure, thereby increasing the efficiency of uncertainty analysis. Because sensitivity analysis of complex models can be highly computationally demanding, it is a focus of current research to help improve efficiency and applicability.

Typically, sensitivity methods are categorized into local (LSA) and global sensitivity analysis (GSA) techniques. Basically, LSA methods analyze sensitivity of model responses around some point in input parameter space (ideally around optimal locations), while GSA methods analyze the variability of model responses across the full parameter space (Figure 14). Figure 14 Illustrates the concept of local and global sensitivity analysis for a model with two parameters. For a model with larger number of parameters the 2D response surface will change to a more than 2-dimension (dimension dependent on the number of parameters) response space. Each black dot represents a combination of parameters used to quantify model response and ultimately determine the sensitivity of model response to each parameter.

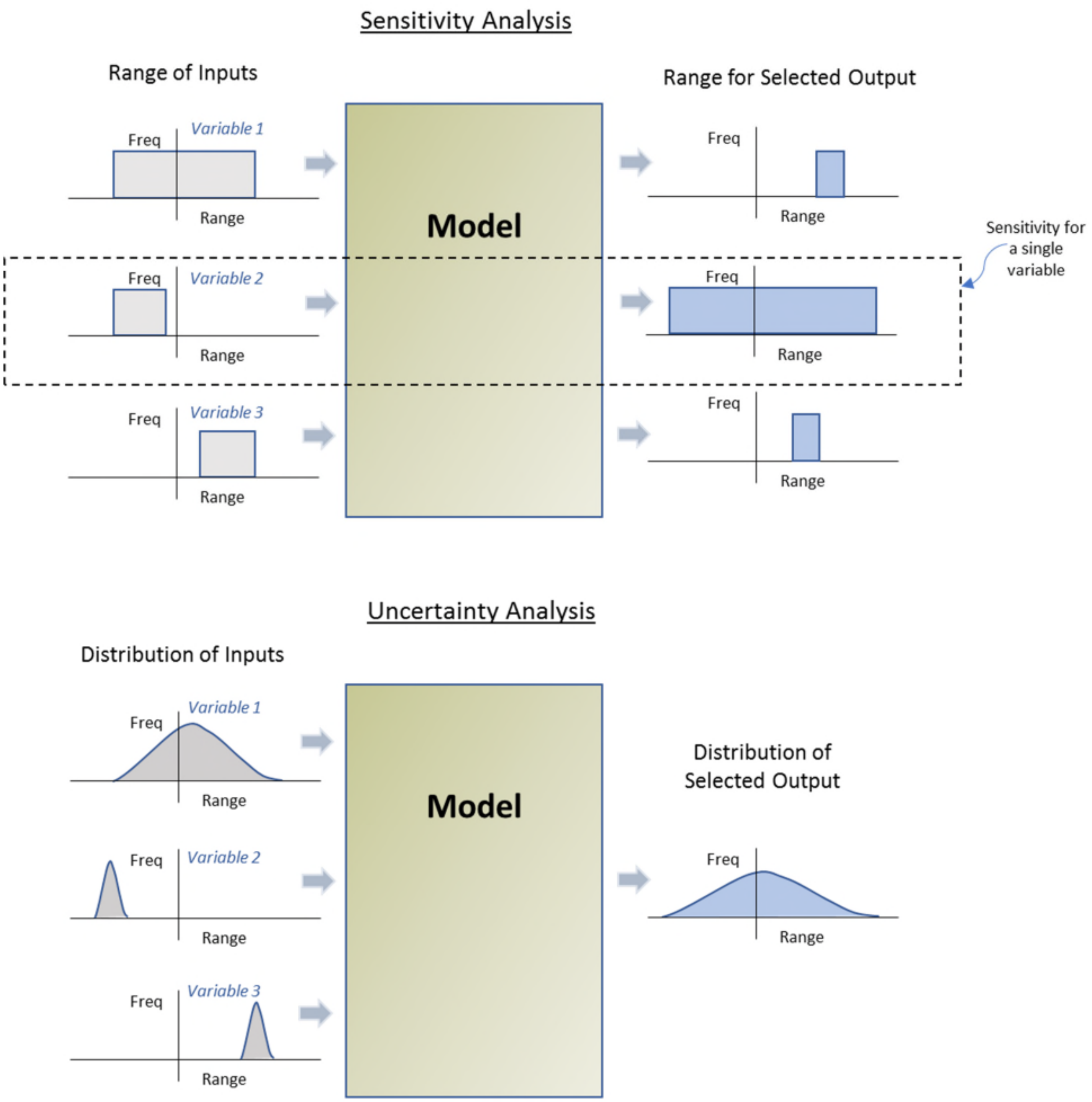


Figure 13. Simplified representation of sensitivity and uncertainty analyses. Inputs in this context may include parameter values, initial conditions and boundary conditions that are used for a single model run. During sensitivity analysis a model is run with a range of values for key inputs and the corresponding range in one or more outputs is evaluated. As part of uncertainty analysis, inputs are assigned ranges in values based on known estimates.

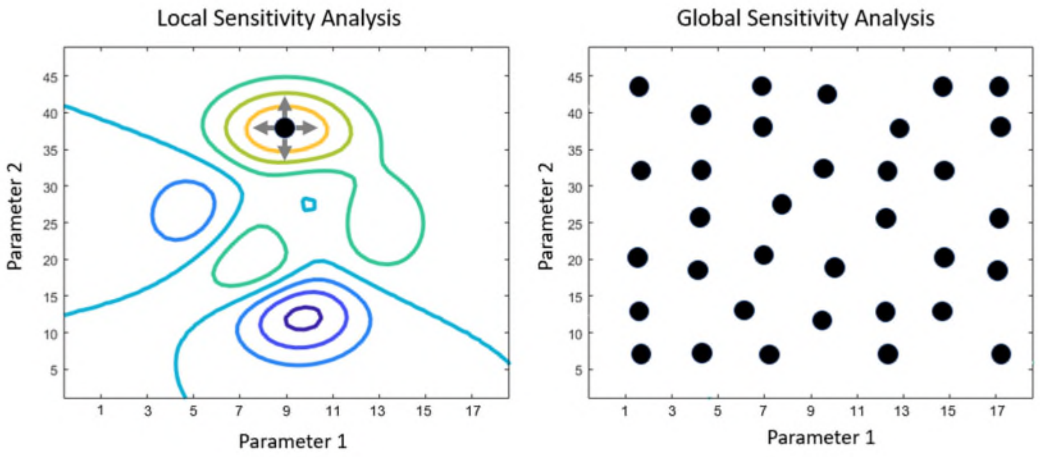


Figure 14. Illustration of the concept of local (left panel) and global sensitivity analysis (right panel) for a model with two parameters. For a model with larger number of parameters the 2D response surface will change to more than 2-dimension (dimension dependent on the number of parameters) response space. Each black dot represents a combination of parameters used to quantify model response and ultimately quantify sensitivity to each model parameter.

LSA is a partial derivative-based method to investigate the response of a small disturbance of each parameter around a specific location in parameter space on model output (Baroni and Tarantola, 2014). A common approach for conducting LSA is the one-factor-at-a-time (OAT) method (Yang, 2011). In OAT, one parameter is changed at each iteration. LSA techniques are appropriate for relatively simple models that show linear responses. Although LSA is computationally efficient and popular, it is not suitable for reducing the dimensionality of complex non-linear environmental models as it disregards the correlation between model parameters, and its results are dependent on location and often there is a lack of knowledge on the suitable location, i.e., the parameter true value (Saltelli et al., 2008).

GSA investigates the effect of variations over the entire prior parameter space on model output (Saltelli et al., 2008; Pianosi et al., 2016). A sensitivity analysis approach that is commonly used with GSA is the “All-at-a-time” (AAT) approach. GSA does not have the limitations associated with LSA, as it does not rely on a pre-known optimal location for parameters. A common approach for GSA is rooted in relating the variance of the model responses to the change in input parameters (variance-based techniques). Variance-based sensitivity methods have shown very promising results. However, the sample size required to achieve reasonably accurate approximations can be rather large, which compromises their applicability to highly complex models. Several methods have been proposed to reduce the required number of model evaluations for approximating the variance-based indices. These include: (i) methods using the Fourier series expansion of the model outputs, such as Fourier Amplitude Sensitivity Test (FAST) for the approximation of the first-order indices, and the extended FAST for the total-order indices; and (ii) methods rooted in application of a model emulator which will be discussed further in proceeding sections.

3.3 Uncertainty Quantification and Propagation

Models, as simplifications of reality, are subject to various forms of uncertainty. In environmental models specifically, these sources of uncertainty include: i) parameters, ii) structure (model conceptualization), iii) initial state variables, iv) configuration and input variables, and v) observation data used for training and testing the model. Further, the nature of uncertainty can be categorized into epistemic uncertainty and aleatory uncertainty or stochastic uncertainty (Walker et al., 2003). Epistemic uncertainties stem from our lack of knowledge and they can be reduced with additional collection of data. In contrast, aleatory uncertainties originate from inherent variability and stochasticity of natural phenomena (e.g., climatic variability). Aleatory uncertainties cannot be reduced by collection of more data. For certain natural phenomena, this means that there is no direct way of getting perfect knowledge. Climate predictions over different time scales are perhaps the most common example of aleatory uncertainty in environmental models. Modeling applications typically include both epistemic and aleatoric uncertainties.

The lack of accounting for uncertainties when applying models may result in biased and unreliable results which will directly affect the decisions made based on the modeling results (Beven and Binley, 1992; Refsgard et al., 2007; Bastin et al., 2013). Various methods have been proposed to address the uncertainties from model parameters (Moradkhani et al., 2005), input data (Kavetski et al., 2003), monitoring data (Harmel and Smith, 2007), and model structure (Ajami et al., 2007) in hydrologic and water quality models.

Uncertainty assessment methods fall under one of two classifications: forward uncertainty propagation and inverse uncertainty quantification. In forward propagation methods, uncertainties in model inputs are propagated to the model outputs. In inverse uncertainty quantification methods, posterior distributions of model parameters are derived based on discrepancies between model simulations and observations and values of likelihood function. Inverse quantification of uncertainty is much more complex than forward propagation of uncertainty, as the modeler is essentially solving the problem in reverse (similar to calibration). However, the method provides essential benefits when modeling as in most cases the uncertainties associated with various model elements (parameters, inputs, etc.) are initially unknown and using an inverse approach, the modeler can estimate the most consequential uncertainties, and select them for further evaluation. Thus, these uncertainties can be propagated to simulations through a forward approach. In most inverse uncertainty quantification applications, the overall modeling uncertainties are quantified as a lumped value as quantifying the uncertainties associated with each model components is very time-consuming and in some cases impossible. Specifically, in highly complex integrated environmental models, decomposition of uncertainty and attributing portions of total uncertainty (total error) to various sources of uncertainty is an extremely challenging task which still is a subject of extensive ongoing research (Bastin et al., 2013).

Bayesian-based methods are among the most commonly used assessment techniques for conducting uncertainty analysis for complex environmental models (Jia et al., 2018). Bayesian uncertainty analysis methods, rooted in Bayes' Theorem, quantify parameter uncertainty by deriving the posterior parameter distribution from a combination of prior parameter distribution and a likelihood function. In most environmental models, specifically more complex models, the analytical solution to derive the explicit functional form of the posterior distribution is infeasible. Hence, sampling is often used to derive the posterior distribution. The Markov Chain Monte Carlo (MCMC) sampling schemes provide efficient algorithms to derive the posterior parameter distribution (Rath et al., 2017; Tasdighi et al., 2018). In this regard, multi-chain MCMC methods have proven superior performance and efficiency in sampling the parameter space and deriving the posterior distributions. Application of multiple Markov chains enhances the efficiency of the search algorithm and reduces the chance of being trapped in local optima (Ter Braak, 2006). Two common multi-chain MCMC algorithms frequently used for environmental models are the DiffeRential Evolution Adaptive Metropolis (DREAM) algorithm (Vrugt, 2016) and the Shuffled Complex Evolution Metropolis (SCEM) algorithm (Duan et al., 1992; Vrugt et al., 2003). While multi chain MCMC algorithms have been employed in conducting uncertainty analysis for various environmental models, their application to integrated model frameworks remain very limited due to computational burden (Tscheickner-Gratl et al., 2019).

Another source of model uncertainty, discussed much less frequently, is related to human decisions regarding process representation in a model, including underlying assumptions and prioritization. Although this human-imposed bias cannot be evaluated quantitively, it should be considered as part of a broader validation and peer-review exercise.

3.4 Frameworks for Model Calibration, Sensitivity, and Uncertainty Analysis

Most model evaluation techniques are built around iterative approaches that entail running the model multiple times. For large models with long run times, iterative approaches can be very challenging. Integrated environmental models pose particular challenging as they are highly parametrized and often have multiple component models that work jointly to generate results. General approaches to address this issue include 1) employing more computational power (more computational capacity from software and hardware), and 2) revising the model evaluation algorithms to be more efficient in exploring the model response space. Often, both approaches are needed for evaluating complex models. A summary of common tools for conducting model evaluation is presented below.

Table 4. Common Tools for Model Calibration, Sensitivity, and Uncertainty Analysis

Tool	Features
PEST	**PEST** (Parameter Estimation and Uncertainty Analysis) was the first model-independent tool of its kind and has gained a large, diverse following in various branches of environmental modeling. While initial versions of PEST only supported Gauss-Marquardt-Levenberg (GML) parameter estimation technique, more recent versions have added numerous features including: sophisticated regularization schemes, globalized GML (GGML) capabilities, implementations of the Shuffled Complex Evolution (SCE) and Covariance Matrix Adaption Evolutionary Strategy (CMA-ES) global search algorithms, post-regression diagnostics, and predictive uncertainty analysis capabilities. PEST is freely available to the public and can be accessed from: http://www.pesthomepage.org/Home.php. Papadopulos & Associates Inc. have recently developed two commercial versions of PEST that benefit from high performance parallel computing and cloud-based computing, making it specifically suited for integrated environmental models. The company uses Microsoft Azure Cloud for running highly parallelized PEST, which substantially reduces the run time and provides essential benefits for sensitivity, calibration, and uncertainty analysis. The service is commercial, and details can be accessed from: https://www.sspa.com/training/pesthp-and-pestcloud
DAKOTA	DAKOTA toolkit links to a variety of well-established model optimization packages and libraries into a model-independent flexible package. The tool's advanced parametric analyses enable design exploration, model calibration, sensitivity analysis, risk analysis, and quantification of uncertainty. DAKOTA users can select from a diverse and ever-expanding suite of parameter estimation algorithms. Available algorithms span the entire range of single solution approaches (i.e., local, global and hybrid), and numerous multi-objective and surrogate-based options. Dakota is open source, with applications spanning environment and climate modeling, computational materials, nuclear power, renewable energy, and many others. DAKOTA can be accessed from: https://dakota.sandia.gov/
OSTRICH	OSTRICH (Optimization Software Toolkit for Research Involving Computational Heuristics) toolkit is a model-independent and multi-algorithm optimization and calibration tool. It can be used for weighted non-linear least-squares calibration of model parameters or for constrained optimization of a set of design variables according to a user-defined objective or cost function (single or multi-objective supported). OSTRICH implements numerous local, global, and hybrid search algorithms, including multi-start GML, Particle Swarm Optimization (PSO), and Dynamically Dimensioned Search (DDS). OSTRICH also contains a module for efficient multi-model calibration, ranking and selection. OSTRICH supports Message Passing Interface (MPI)-based parallel processing on both Windows and Linux machines. The parallel version of OSTRICH is called OstrichMPI. OSTRICH is available for free to the public and can be accessed from: http://www.eng.buffalo.edu/~lsmatott/Ostrich/OstrichMain.html
UCODE	UCODE is a model-independent toolkit for conducting model sensitivity, calibration, and uncertainty analysis. UCODE implements the gradient-based GML, non-linear regression algorithm, and calculates numerous post-regression statistics. UCODE input/output files follow the Joint Universal Parameter IdenTification and Evaluation of Reliability (JUPITER). Application Programming Interface (API) specification and can be utilized directly by similarly compliant uncertainty assessment programs, such as Multi-Model Averaging (MMA), a tool for multi-model analysis. UCODE is available for free to the public and can be accessed from: https://igwmc.mines.edu/ucode-2/
SimLab	SimLab is a software designed for Monte Carlo (MC)-based uncertainty and sensitivity analysis (SA). Various sampling procedures are used for conducting MC. Among those are: random sampling, stratified sampling (including Latin Hypercube Sampling), and quasi-random sampling. SimLab supports screening-level methods along with several variance-decomposition algorithms, including Fourier Amplitude Sensitivity Testing (FAST), extended FAST (eFAST), and method of Sobol for Global Sensitivity Analysis. SimLab also supports numerous regression and correlation-based SA techniques. SimLab toolkit is free to download for the public and can be accessed from: https://ec.europa.eu/jrc/en/samo/simlab

Tool	Features
GLUE	GLUE (Generalized Likelihood Uncertainty Estimation), utilizes a Monte Carlo importance-sampling procedure to locate behavioral parameter sets and estimate parameter distributions. The mGLUE tool enhances GLUE efficiency utilizing an artificial neural network as a surrogate for model evaluation. The network is trained using the results of a genetic algorithm (GA), in which “niching” reduces bias in the subsequent surrogate-based GLUE samples. GLUE has been applied to a wide variety of fields including rainfall-runoff modelling, flood inundation, water quality modelling, sediment transport, recharge and groundwater modelling, vegetation growth models, aphid populations, forest fire and tree death modelling. There are multiple versions of GLUE in different programming languages available to the public for free. A Python version can be accessed from: https://pypi.org/project/pyGLUE/

3.5 Novel Approaches for Confronting Computational Burden of Sensitivity and Uncertainty Analysis for Complex Model Frameworks

Conducting sensitivity and uncertainty analysis for complex model frameworks is often challenging due to the high computational cost of running them. Indeed, continually exercising the simulator to carry out tasks such as sensitivity analysis, uncertainty analysis, and parameter estimation is often infeasible (Baustert et al., 2018). Modelers are then faced with only a limited number of model realizations for conducting their analysis. A common approach sometimes used to overcome this pitfall is application of model emulators. Application of cloud-based platforms and parallel computing are also becoming possible with more computational power from supercomputers and clusters.

Emulators. Emulators basically represent the input/output relationships in a model with a statistical surrogate to reduce the computational cost of model exploration. In this approach, the computer model is viewed as a black box, and constructing the emulator can be thought of as a type of response-surface modeling exercise (Box and Draper, 1986). The aim is to establish an approximation to the input-output map of the model using a limited number of complex model runs. For instance, an emulator can be used to stand in place of a complex computer model when conducting sensitivity or uncertainty analysis. Of course, as with any approximation, there is a reduction in the accuracy of the estimates obtained, and the trade-off between accuracy and cost needs to be considered by the modeler.

There are two commonly used emulators for mimicking complex environmental models (National Research Council, 2012). The first type attempts to approximate the dependence of the computer model outputs on the inputs. In this case, the uncertainty comes from not having observed the full range of model outputs or from the fact that another model is used in place of the costly computational model of interest. These emulators include: (i) regression models; (ii) Gaussian process (GP) interpolators and Lagrangian interpolations of the model output; and (iii) reduced-order models. The second type of emulators are similar, with the additional considerations that the input parameters are now themselves uncertain. So, the aim is to emulate the distribution of outputs, or a feature thereof, under a prespecified distribution of inputs. Statistical sampling of various types (e.g., Monte Carlo sampling) can be an effective tool for mapping uncertainty in input parameters to uncertainty in output parameters (McKay et al., 1979). In its most fundamental form, sampling does not retain the functional dependence of output on input, but rather produces quantities that have been averaged simultaneously over all input parameters. Alternatively, approaches such as polynomial chaos attempt to leverage mathematical structure to achieve more efficient estimates of quantities of interest (National Research Council, 2012).

Greater Computational Resources. Another approach for confronting the high computational demand of complex models and integrated model frameworks is cloud-computing. Cloud-computing is the on-demand availability of computer system resources, especially data storage and computing power, without direct active management by the user. The term is generally used to describe data centers available to many users over the Internet. Large clouds, predominant today, often have functions distributed over multiple locations from central servers (Chen et al., 2018). Clouds provide vast amounts of computational power in a very short amount of time which can be used to accelerate the execution of sensitivity and uncertainty analysis runs for complex models. Parallel computing resources may also be deployed in such cases and have been applied for calibration of certain models in the Delta.

3.6 Alternative Models

For key problems, it is possible (and even preferable) to consider different models to evaluate scenarios. Where multiple models are available—as is case with models for simulating hydrodynamics, watersheds, and groundwater—ranging in theoretical formulation, complexity, availability and accessibility, it is worthwhile to perform comparative studies and evaluate model performance under different conditions. This may be especially beneficial when the modeling effort is being used to support a major, consequential decision. Differences across models provide insight into potential sources of error or inadequate representation in the conceptual model. In some cases, more complex models (e.g., three-dimensional fluid flow models versus two- or one-dimensional models) may provide more nuanced results, but in other cases simpler models may be easier to communicate with decision-makers and stakeholders. Indeed, where possible, there is a benefit to supporting a hierarchy of models with different levels of complexity to better communicate with different users and potentially to allow different levels of integration across models.

There is also a potential downside to implementing alternative models for the same problem, in terms of diluted resources, disagreements between model experts on narrow model formulation issues, and confusion and lack of confidence on the part of stakeholders. Consequently, alternative models should not be considered as a general solution, but considered for major, complex problems where alternate theoretical formulations are possible. Perhaps the most common examples of alternative models used in the Delta are those for climate change projections (where a suite of models has been recommended for use in climate studies by Cal-Adapt, www.cal-adapt.org) and hydrodynamics and salinity transport in the estuary (the SCHISM, UnTRIM, RMA and Delft3D models, as described in Memo 1).

3.7 Post Audit: Compare Model Results to Future Data Being Collected

For models that are used to make near-term forecasts, but also for longer-term predictions, it is important to revisit model outcomes and to compare field observations with previously made model predictions. This process is termed a post-audit, and its importance has been highlighted in other modeling guidance as well, notably, the *Guidance on the Development, Evaluation, and Application of Environmental Models* (USEPA, 2008). For major models that are often in use for a decade or longer, and where supporting observed data continue to be collected, a post-audit is not very difficult to implement. A post-audit can provide insight on conditions under which model performance was acceptable and in line with prior calibration history, thus providing credibility to the model and related modeling studies. A post-audit may also result in the opposite outcome. Under conditions where model performance was poorer than expected, the post-audit provides an excellent opportunity to revisit the fundamental conceptual model and/or the model calibration. Indeed, a post-audit can provide an excellent basis for future model improvements.

3.8 Compatibility with Existing Data Exchange Standards

Even where models are not used within an integrated framework, a model's outputs are rarely used in isolation. Therefore, it would be beneficial to have common frameworks for exchanging data between different models and data processing/visualization tools. Agreed upon standards exist in specific disciplines, such as the DSS standard in hydrology and multiple standards for geospatial and other model data exchange (see Memo 3 for specific frameworks). However, over the broad range of disciplines considered in this work, there are major differences in how data are represented in space and time, and how data are stored and transferred between models. Thus, for most integrated modeling studies, it is not unusual for a large amount of analyst time to be spent in moving data across formats. Where these data transfer processes cannot be automated, there are clear limitations to model integration. To address this practical challenge, focused efforts and collaboration across developers in different areas are needed to provide tools that enable efficient exchange across a variety of models.

3.9 Summary

This chapter describes a set of actions that can help strengthen models, but in almost all cases require substantial additional resources to implement. Indeed, in some cases, the methodologies are the subject of ongoing development and research, and their use may bring forth additional challenges. For this reason, these actions are recommended for use in major modeling studies tied to large societally consequential decisions. These additional actions include: i) peer review of model studies at various stages of implementation; ii) sensitivity analysis of models to key drivers including adjustable parameters and boundary conditions; iii) uncertainty analysis of model studies; iv) consideration of novel approaches to meet the sensitivity and uncertainty analysis needs of complex models and model frameworks; v) consideration of alternative models for model studies, where available; vi) performance of post-audits, i.e., review and evaluate historical model predictions in light of new field observations; and vii) development and compatibility with exchange standards to enable data sharing across models.

4 Broader Issues in Modeling

Because models are employed within a larger decision-making framework in the Delta, it is helpful for modelers and model users to think beyond the strictly numerical and even conceptual frameworks addressed in Chapters 2 and 3. Even when many of the steps identified in the preceding chapters are implemented, a model study may not achieve the support and credibility that it needs to be successful. This document focuses on a set of issues that go beyond the technical aspects of model development and testing, to develop effective products that are useful for decision-makers and stakeholders. Attention to these issues will allow model results to be accessible to a broader audience, including the scientific and educational communities in the region. This chapter provides some general insights based on our experience on working with complex, multi-faceted modeling problems in the Delta region.

4.1 Communication Strategy for Model Study

At the inception phase of a modeling study, during project execution, and at its completion, it is important to think through the overall communication approach with model sponsors and stakeholders, and the larger community of technical experts and the general public. Some of these steps are outlined elsewhere in this document and are summarized below:

- **At project inception:**
  - Outline expected use and purpose of the model to sponsors and stakeholders; run through a model evaluation checklist.
  - Define conceptual model and communicate key processes and unknowns at the start of the study. For major studies, solicit peer review and feedback on the conceptual model.

  - Check on the availability of sufficient data that is legally and scientifically defendable. Implement a monitoring program if necessary.

- **During project execution:**
  - Solicit peer review at key interim steps of the modeling process, focusing on items such as data available, assumptions, and methodology.
  - Update stakeholders on progress and changes in approach from initial plan.
- **At project completion:**
  - Perform peer review of results, including calibration and evaluation, as well as further evaluation such as sensitivity and uncertainty analysis.
  - Prepare documentation on modeling study.
  - Review and update conceptual model and share with stakeholders. Conduct workshop to share results.
  - Develop summary sheets for a broad, general audience. The summary should describe the problem and outcomes with minimal technical jargon.

4.2 Consideration of Model Bias

Modeling is a human activity and is inherently subject to bias. It is essential to acknowledge this bias and attempt to minimize it. This general concern has been addressed in other published guidance as well (Glynn, 2008). Several biases worthy of attention in modeling of complex systems are listed below. Not all of these biases can be addressed through the steps listed in Chapters 2 and 3, and it is important that modelers and model users keep these in mind when considering the practical consequences of a modeling study:

- **Confirmation bias.** Modelers often focus on and highlight observations that confirm a pre-existing conceptual model and are less willing to seek data that will counter the model. This is also termed as "group-think" and is associated with a resistance to new approaches for analysis.
- **Temporal insensitivity bias.** Decision makers may have more interest in predictions that are one to two generations in the future than in the distant future.
- **Steady-state bias.** Modeling approaches often assume constancy in conditions or assume that known variability will continue into the future. In the current literature, this assumption of stationarity is questioned most often by the increasing realization of climate change impacts on aquatic ecosystems. However, this bias is not limited to climate change alone and could involve virtually any social or economic system.
- **Disciplinary bias.** Modelers tend to focus on topics that they know most about even in the case of a framework where multiple models are being integrated.
- **Separation between man and nature.** Mechanistic models of natural systems often do not explicitly account for changes in human behavior during the period of simulation. Thus, regulatory actions are implemented in models as fixed drivers, but not as variables themselves. There is growing interest in developing models that represent as dynamic actors, with responses changes as natural conditions change.

4.3 Results Gauged to Different Audiences

Model findings will be used by and will need to satisfy the information needs of different audiences, from technical specialists to members of the general public. Therefore, it is important that modelers are also engaged at different levels of this process such that the right information is transferred to each audience. Furthermore, audiences may weigh in on a modeling study during various phases of the project. Considerations for different audiences at project inception and completion are described below:

- **At project inception:**
  - **Technical specialists.** Such audiences will need to understand why the modeling is needed and the approach to be used. Some of the items in the project inception checklist may serve to aid this goal. It is also important to convey the novelty of a modeling exercise, and how it extends current thinking.
  - **Stakeholders.** Such audiences will need to know the specific answers to be obtained through the modeling and whether similar answers can be developed without modeling. They will need to know the costs, time frames, and major unknowns. Modelers should use this opportunity to highlight known and potential uncertainties, and how this might affect the outcome of the findings. Conceptual models can be used as a tool to highlight the areas of focus of the modeling exercise. In communication with stakeholders, it is important to have clear definitions, and minimize use of jargon as much as possible.
- **Near completion:**
  - **Technical specialists.** Such audiences will expect to see many of the technical steps described in Chapters 2 and 3, such the basis of the model, specific assumptions used, the results of testing and evaluation of uncertainty. It is also important to convey the novelty of a modeling exercise, and how it extends current thinking. The presentation of key model studies as peer-reviewed publications provides additional credibility and also provides archival benefits for a modeling exercise. Finally, audiences may want to understand next steps or the long-term plan for the study (additional modeling or data collection, etc.).
  - **Stakeholders.** Such audiences need a high level overview of key findings that can be quickly understood across broad range of people, expertise and experience. Good results are simple and memorable and tell the key elements of a story in a compact manner. A new or updated conceptual model is a good summary of the overall exercise. Additional graphical resources, beyond the conceptual model, may also be developed to help readers understand important findings. It is helpful to describe what was achieved through the modeling and what remains unknown.

4.4 Building Stakeholder Engagement and Trust

With the growing adoption of models in support of socially consequential decisions, it is becoming increasingly important that models be considered credible by the stakeholder community. This is a feature of modeling applications worldwide, and a trend away from technocratic decision-making to a more participatory framework (Voinov and Bosquet, 2010; Voinov et al. 2016; Parrott et al., 2017). Because of the complex and well-established interaction of many environmental processes in the Delta (e.g., water supplies and ecosystems), stakeholders for a modeling exercise include a range of participants, from government agencies at different levels (local, state, and federal) with different areas of focus, as well as non-governmental and individual participants with different levels of expertise and different interests. Decision-makers are often faced with situations where some stakeholders are not convinced of the benefits or the credibility of a model used to support a decision; thus, social as well as scientific credibility should be pursued for a model and related studies. Systematic engagement with stakeholders on various aspects of a modeling exercise—such as modeling purpose, conceptual model, calibration and evaluation, and peer review (as described above) – may provide such credibility and support over the long-term.

Voinov et al. (2016) present a comprehensive review of stakeholder engagement processes and tools in recent modeling efforts. They observe a demand for greater citizen engagement in planning and policy decisions (of which environmental modeling is a part) and note that tools and processes for sharing such data are rapidly evolving. They also note rapid improvements in graphical tools and internet-based tools to display information and new social media formats for exchange of information. Citizens, less in awe of the mystique of models (or experts in general), are more able to participate and contribute to modeling processes through such mechanisms. These drivers, and a need for meeting greater stakeholder expectations, will (over time) change how models are packaged and presented. The need for transparency in formulation, assumptions, and presentation of inputs and outputs will continue to grow.

4.5 Model Utility and Friendliness

As a model becomes used more frequently and/or used within a larger stakeholder community, the justification for resources to support documentation and solid practices becomes stronger and more obvious. Model accessibility, including utility and friendliness, has an influence on how frequently and/or broadly a model is used. Below are design considerations for increasing model accessibility:

User-friendliness and lower barriers to entry. All models require certain levels of expertise and skills. High barriers to entry and poor user-friendliness (e.g. poor documentation, arcane technology) limit the pool of people that will utilize and leverage the model. Making models more user-friendly broadens that pool and brings greater value to society. This consideration has been a factor in some of the more successful technical models in use today.

Models to provide initial scoping results. Simple and easy-to-build models can be useful as tools to quickly assess a situation with early results that provide initial scoping information. These scoping results can provide a foundation for considering more complex models in subsequent analyses. These types of models can find broad audiences for smaller problems akin to the broad use of spreadsheet analyses today. Models for initial scoping may be helpful to engage stakeholders and are preferable to a situation where even initial results are dependent on a model that takes a long time to develop.

Fun, Entertaining and Accessible. Technology becomes more attractive when it is fun, entertaining and accessible. Graphs that can be displayed on phones or tools that can be readily accessed on the web are thus more likely to be used. Augmented reality (AR) tools that present model results in the context of the real world will allow greater engagement and understanding of outcomes. An example of this is the model SimBasin, which overlays the WEAP water resources modeling software. SimBasin uses a model of an actual basin and is designed to facilitate communications and engagement between stakeholders and scientist when considering different policy impacts.

4.6 Model Sustainability over Long Time Horizons

Models represent large intellectual and financial investments, but in most instances, their long-term viability is unknown. Models are often developed to serve a specific need, and access to these models and related analyses rapidly diminishes over time. For key foundational models and related efforts, long term sustainability should be addressed early in its life cycle to make best use of the investments being made. This life cycle planning should identify the responsibilities, accountabilities, and resources needed to support a model over the long term, potentially over decades. This life cycle planning should also contemplate development of new versions and ongoing model evaluation. Because of the resources and long-term commitments required to sustain a model over time, this is an issue that extends beyond the modelers and should be brought to the attention of decision and policy makers early in the process.

4.7 Summary

The technical strength of a model can be established through the steps presented in Chapters 2 and 3. Nonetheless, there remain several non-technical issues that a modeler should address to meet the broader goals of a modeling exercise. These non-technical issues are discussed in this chapter and include: development of a communication strategy for a modeling study, consideration of bias in many aspects of the model formulation, presentation of results across many audiences, building trust across the community that will be using the model results, overall user-friendliness of the modeling framework, and practices for sustaining the usefulness of a model over a long-term horizon.

5 Encouraging Adoption of Best Practices

Many of the concepts identified in the preceding chapters are perhaps known to most modelers but are not widely adopted. This may be due to time and resource limitations associated with virtually all modeling studies; this may also be due to the lack of specific expectations in the broader community of modelers and model users. Thus, model users may not know what specific and reasonable requests to make of modelers to guide a model study toward greater credibility and usefulness.

To encourage adoption of the best practices identified in this work, we provide three relatively compact summary sheets. The purpose of the first sheet (Table 5), designed as a checklist to be employed at inception of a modeling effort, is to enable various participants to agree on the basic features of the work to be done. The purpose of the second sheet (Table 6) is to evaluate and score a modeling exercise upon completion. A final sheet (Table 7) is designed to help evaluate the long-term sustainability of a modeling framework.

The first sheet is designed with Yes/No responses, although additional narrative information can be provided. While there are no correct answers associated with the model study pre-audit, the questions are designed to flag issues that may need to be resolved before significant modeling study resources have been expended.

The second sheet contains a list of questions that may be answered with narrative responses or with numerical scores. If the numerical scoring approach is used, a model study with a higher score is more desirable. A numerical scoring approach may be useful for comparing multiple model studies that employ the same type of domain modeling. However, this approach is of limited value when a unique or one-of-a-kind model study is to be evaluated. The questions provided in these sheets are offered as starting points to be modified as needed for specific agencies or applications. However, we expect many of the essential items will apply to most modeling studies.

The third sheet is focused not on modeling *per se*, but on questions that help evaluate the long-term sustainability of a model framework. It is not intended to evaluate a single study, but to assess whether the framework used in one or more studies is well supported into the future.

Table 5. Model Study Initial Appraisal

Item	Description	Response
1	Do we know how the model results will be used?	Yes/No
2	Is the model to be used defined?	Yes/No
3	Has a conceptual model been developed?	Yes/No
4	Have the criteria for selecting the model been defined?	Yes/No
5	Is an existing model going to be modified?	Yes/No
6	Is a new model to be developed?	Yes/No
7	Are the time frames known for initial model development, calibration, testing, and review?	Yes/No
8	Are data associated with intended model inputs available?	Yes/No
9	Does the model need calibration?	Yes/No
10	Are data associated with intended model outputs available (to support model calibration)?	Yes/No
11	Are time frames of the input and output data known and consistent with one another?	Yes/No
12	Are the errors in data measurements known?	Yes/No
13	Is the level of error in the expected results known?	Yes/No
14	Are the model stakeholders known?	Yes/No
15	Will stakeholders be part of the modeling process?	Yes/No
15	Have users of the model output met together?	Yes/No
16	Will documentation be prepared upon completion of the model?	Yes/No
17	Will the information embedded in questions 1-15 be used to prepare a memo describing the model's purpose?	Yes/No

Table 6. Model Study Post-Completion Appraisal

Item	Description	Response (Numeric Score or narrative)
1	Is the model a new formulation or the application of an existing code? If a new formulation, what has been done to test and verify the code?	
2	Has a conceptual model been developed for this effort and has it been updated following completion?	
3	Are observed data used in the modeling exercise (input and output data) documented and available for review?	
4	Has the calibration approach been described?	
5	Has the model performance following calibration been adequately evaluated using test data?	
6	Has the sensitivity of major variables been evaluated?	
7	Has model output uncertainty been evaluated?	
8	Were any novel approaches used to evaluate the sensitivity and uncertainty of the model response to inputs?	
9	Were the model results compared and contrasted with other models (if available)?	
10	Does the model study documentation adequately explain the approach, assumptions, and findings?	
11	Was a peer review performed and responded to?	
12	What were the stakeholder’s reactions to the model results?	
13	Are the model summary documents easily understandable by a variety of audiences?	

Table 7. Model Framework Life Cycle Evaluation

Item	Description	Narrative Response
1	Are all source codes and supporting files stored in a single location and archived in a manner that enables future access?	
2	Are the source codes documented, even if this documentation is not in the public domain?	
3	Is the model development dependent on a single individual? What is the long-term transition plan for the expertise in this model?	
4	Is the model framework applied by a community or by a single team? Is there a mechanism to share knowledge about the model application over time, such as a virtual community, trainings, etc.?	
5	Is there a defined plan for making updates to the model framework?	
6	For a public-domain model framework, is there a funding mechanism to support staff that would work on the model?	
7	For a proprietary model framework, what is the mechanism to support the code development over the long-term?	

5. Encouraging Adoption of Best Practices

6 References

Ajami, N.K., Duan, Q., Sorooshian, S., 2007. An integrated hydrologic Bayesian multimodel combination framework: confronting input, parameter, and model structural uncertainty in hydrologic prediction. Water Resour. Res. 43 (1), W01403.

Baroni, G., Tarantola, S., 2014. A general probabilistic framework for uncertainty and global sensitivity analysis of deterministic models: a hydrological case study. Environmental Modelling and Software, 51, 26-34.

Bastin, L., Cornford, D., Jones, R., Heuvelink, G.B.M., Stasch, C., Nativi, S., Mazzetti, P., and Williams, M., 2013. Managing uncertainty in integrated environmental modeling: The UncetWeb framework. Environmental Modelling and Software, 39, 116-134.

Baustert, P., Othoniel, B., Rugani, B., Leopold, U., 2018. Uncertainty analysis in integrated environmental models for ecosystem service assessments: Frameworks, challenges and gaps. Ecosystem Services, 33(2018) 110-123.

Beven, K.J., Binley, A.M., 1992. The future of distributed models: model calibration and uncertainty prediction. Hydrol. Process. 6, 279–298.

Beven, K.J., Freer, J., 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. J. Hydrol., 249(1-4), 11-29.

Blaise, B., 2019. Lawrence Livermore National Laboratory, Introduction to Parallel Computing, available at: https://computing.llnl.gov/tutorials/parallel_comp/

Blöschl, G., Sivapalan, M., 1995. Scale issues in hydrological modelling: A review, Hydrol. Process., 9, 251-290.

Box, G.E.P., and Draper, N.R., 1986. Empirical model-building and response surface. John Wiley & Sons, Inc. New York, NY, USA.

California Water & Environmental Modeling Forum (formerly Bay-Delta Modeling Forum), 2000. Protocols for Water and Environmental Modeling.

Chen, C., Che, D., Yan, Y., Zhang, G., Zhou, Q., and Zhou, R., 2018. Integration of numerical model and cloud computing. Future Generation Computer Systems, 79(1) 3960407.

Crout, N., Kokkonen, T., Jakeman, A.J., Norton, J.P., Newham, L.T.H., Anderson, R., Assaf, H., Croke, B.F.W., Gaber, N., Gibbons, J. and Holzworth, D., 2008. Chapter 2: Good modelling practice. Developments in Integrated Environmental Assessment, 3, pp.15-31.)

Daniel, E.B., Camp, J.V., LeBoeuf, E.J., Penrod, J.R., Dobbins, J.P. and Abkowitz, M.D., 2011. Watershed modeling and its applications: A state-of-the-art review. The Open Hydrology Journal, 5(1).

Dawson, C., Fringer, O., He, R., Ralston, D., Zhang, J., 2019. Final Report on NSF Workshop on the future of coastal and estuarine modeling.

Doherty, J. and Johnston, J.M., 2003. Methodologies for calibration and predictive analysis of a watershed model 1. JAWRA Journal of the American Water Resources Association, 39(2), pp.251-265.

Doherty, J., 2015. Calibration and Uncertainty Analysis for Complex Environmental Models. Watermark Numerical Computing, Brisbane, Australia. ISBN: 978-0-9943786-0-6.

Doherty, J.E. and Hunt, R.J., 2010. Approaches to highly parameterized inversion: a guide to using PEST for groundwater-model calibration (p. 2010). US Department of the Interior, US Geological Survey.

Duan, Q., Sorooshian, S., and Gupta, V.K.. 1992, Effective and efficient global optimisation for conceptual rainfall-runoff models. Water Resources Research, 28 (4) (1992) 1015-1031.

Gaber, N., Foley, G., Pascual, P., Stiber, N., Sunderland, E., Cope, B. and Saleem, Z., 2009. Guidance on the development, evaluation, and application of environmental models. Report, Council for Regulatory Environmental Modeling.

Ganju, N.K., Brush, M.J., Rashleigh, B., Aretxabaleta, A.L., Del Barrio, P., Grear, J.S., Harris, L.A., Lake, S.J., McCardell, G., O'Donnell, J. and Ralston, D.K., 2016. Progress and challenges in coupled hydrodynamic-ecological estuarine modeling. Estuaries and coasts, 39(2), pp.311-332.

Glynn, P.D., 2017. Integrated Environmental Modelling: human decisions, human challenges. Geological Society, London, Special Publications, 408(1), pp.161-182.

Grayson, R. Blöschl, G., 2000. Spatial Patterns in Catchment Hydrology: Observations and Modelling, pp 404. Cambridge University Press, United Kingdom.

Gupta, H. V., Sorooshian, S. and Yapo, P.O., 1999. Status of automatic calibration for hydrologic models: Comparison with multilevel expert calibration. J. Hydrologic Eng. 4(2): 135-143.

Harmel, R.D., Smith, P.K., 2007. Consideration of measurement uncertainty in the evaluation of goodness-of-fit in hydrologic and water quality modeling. J. Hydrol. 337 (3–4), 326–336.

Hawking, S.W., 1988, A brief history of time: From the big bang to black holes. Bantam Books, New York, 1988.

Hudson, R.J., Gherini, S.A., Watras, C.J. and Porcella, D.B., 1994. Modeling the biogeochemical cycle of mercury in lakes: The mercury cycling model (MCM) and its application to the MTL study lakes. Mercury pollution: integration and synthesis, pp.473-523.

Hutton, P.H., Rath, J., Chen, L., Ungs, M.L., and Roy, S.B., 2015. Nine Decades of Salinity Observations in the San Francisco Bay and Delta: Modeling and Trend Evaluation, Journal of Water Resources Planning and Management, DOI: 10.1061 / (ASCE) WR.1943-5452.0000617.

Interagency Ecological Program (IEP), 2015. An updated conceptual model of Delta Smelt biology: Our evolving understanding of an estuarine fish. Available at: https://water.ca.gov/LegacyFiles/iep/docs/Delta_Smelt_MAST_Synthesis_Report_January%202015.pdf.

Jia, H., Xu, T., Liang, S., Zhao, P., Xu, C., 2018, Bayesian framework of parameter sensitivity, uncertainty, and identifiability analysis in complex water quality models. Environmental Modelling and Software, 104(2018) 13-26.

Kavetski, D., Franks, S.W., Kuczera, G., 2003. Confronting input uncertainty in environmental modeling. In: Duan, Q., Gupta, H.V., Sorooshian, S., Rousseau, A.N., Turcotte, R. (Eds.), Calibration of Watershed Models. Water Sci. Appl. Ser. AGU, Washington, D.C., pp. 49–68.

Konikow, Leonard and Bredehoeft, John D., 1992, Ground-water models cannot be validated. Advances in Water Resources 15, 75 – 83

Madsen, H., 2003. Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives. Advances in water resources, 26(2), pp.205-216.

McKay, M.D., Beckman, R.J., and Conover, W.J., 1979, Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. Technometrics, 21:2, 239-245.

Moonen, P. and Allegrini, J., 2015. Employing statistical model emulation as a surrogate for CFD. Environmental Modelling and Software, (72) 77-91.

Moradkhani, H., Hsu, K.L., Gupta, H.V., Sorooshian, S., 2005. Uncertainty assessment of hydrologic model states and parameters: sequential data assimilation using the particle filter. Water Resour. Res. 41, W05012.

Nash, J. E., and Sutcliffe, J.V., 1970. River flow forecasting through conceptual models: Part 1. A discussion of principles. J. Hydrology 10(3): 282-290.

National Research Council (NRC), 2012. Assessing the reliability of complex models: mathematical and statistical foundations of verification, validation, and uncertainty quantification. National Academies Press.

National Research Council, 2007. Models in environmental regulatory decision making. National Academies Press.

Oreskes, N. and Belitz, K., 2001. Philosophical issues in model assessment, in (Anderson, M.G. and Bates, Paul, O., eds.) Model Validation, Perspectives in Hydrological Science, Wiley.

Pianosi, F., Beven, K., Freer, J., Hall, J.W., Rougier, J., Stephenson, D.B., Wagener, T., 2016. Sensitivity analysis of environmental models: a systematic review with practical workflow. Environmental Modelling and Software, 79, 214-232.

Rath, J.S., Hutton, P.H., Chen, L. and Roy, S.B., 2017. A Hybrid Empirical-Bayesian Artificial Neural Network Model of Salinity in the San Francisco Bay-Delta Estuary, Environmental Modeling and Software, 93, 193-208, DOI: 10.1016 / j.envsoft.2017.03.022.

Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., Timmer, J., 2009. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformatics. 25 (15), 1923-1929.

Refsgard, J.C., van der Sluijs, J.P., Hojberg, A.L., Vanrolleghem, P.A., 2007. Uncertainty in the environmental modelling process – a framework and guidance. Environmental Modelling and Software, 22(2007) 1543-1556.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., Tarantola, S., 2008. Global Sensitivity Analysis: The Primer. John Wiley & Sons Ltd, England.

Santhi, C, Arnold, J. G., Williams, J.R., Dugas, W.A., Srinivasan, R., and Hauck, L.M., 2001. Validation of the SWAT model on a large river basin with point and nonpoint sources. J. American Water Resources Assoc. 37(5): 1169-1188.

Schellart, A.N.A., Tait, S.J., Ashley, R.M., 2010. Towards quantification of uncertainty in predicting water quality failures in integrated catchment model studies, Water Res. 44, 3893-3904.

Shoemaker, L.. Lahlou, M., Bryer, M., Kumar, D., and Kratt, K., 1997. Compendium of tools for watershed assessment and TMDL development. USEPA 841-B-97-006.

Shucksmith, J., Heuvelink, G.B.M. and Tait, S., 2019. Recent insights on uncertainties present in integrated catchment water quality modelling. Water Research, 150 (1) 368-379.

Solomatine, D.P., Dibike, Y.B. and Kukuric, N., 1999. Automatic calibration of groundwater models using global optimization techniques. Hydrological Sciences Journal, 44(6), pp.879-894.

Tasdighi, A., Arabi, M., Harmel, D., 2018, A probabilistic appraisal of rainfall-runoff modeling approaches within SWAT in mixed land use watersheds, Journal of Hydrology. 564, 476-489.

Ter Braak, C. J. F., 2006, A Markov chain Monte Carlo version of the genetic algorithm differential evolution: Easy Bayesian computing for real parameter space, Statistical Computing, 16(3), 239–249.

Tetra Tech, 2006. Conceptual Model for Nutrients in the Central Valley and Sacramento-San Joaquin Delta, Technical Report prepared for Central Valley Regional Water Quality Control Board. Available at: https://www.waterboards.ca.gov/centralvalley/water_issues/drinking_water_policy/final_nutrient_report_lowres.pdf

Thiemann, M., Trosset, M., Gupta, H. and Sorooshian, S., 2001. Bayesian recursive parameter estimation for hydrologic models. Water Resources Research, 37(10), pp.2521-2535.

Trowbridge, P.R., Deas, M., Ateljevich, E., Danner, E., Domagalski, J., Enright, C., Fleenor, W., Foe, C., Guerin, M., Senn, D., and Thompson, L.., 2016 Recommendations for a Modeling Framework to Answer Nutrient Management Questions in the Sacramento-San Joaquin Delta. Report prepared for the Central Valley Regional Water Quality Control Board.

Tscheickner-Gratl, F., Bellos, V., Schellart, A., Moreno-Rodenas, A., Muthusamy, M., Langeveld, J., Clemens, F.H.L.R., Benedeti, L., Rico-Ramirez, M.A., Fernandes de Carvalho, R., Breuer, L., Shucksmith, J., Heuvelink, G.B.M., Tait, S., 2019. Recent insights on uncertainties present in integrated catchment water quality modelling, Water Res., 150, 368-379.

U.S. Environmental Protection Agency, 2002. Guidance for Quality Assurance Project Plans for Modeling EPA QA/G-5M, on the Internet at: https://www.epa.gov/quality/guidance-quality-assurance-project-plans-modeling-epa-qag-5m

van Vliet, J., Bregt, A.K., Brown, D.G., van Delden, H., Heckbert, S. and Verburg, P.H., 2016. A review of current calibration and validation practices in land-change modeling. Environmental Modelling & Software, 82, pp.174-182.

Voinov, A. and Bousquet, F., 2010. Modelling with stakeholders. Environmental Modelling & Software, 25(11), pp.1268-1281.

Voinov, A., Kolagani, N., McCall, M.K., Glynn, P.D., Kragt, M.E., Ostermann, F.O., Pierce, S.A. and Ramu, P., 2016. Modelling with stakeholders–next generation. Environmental Modelling & Software, 77, pp.196-220.

Vrugt, J. A., Gupta, H. V., Bouten, W., & Sorooshian, S., 2003. A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. Water Resources Research, 39(8).

Vrugt, J., 2016. Markov chain Monte Carlo simulation using the DREAM software package: theory, concepts, and MATLAB implementation. Environmental Modelling and Software, 75 (2016) 273-316.

Wang, H., Fu, X., Wang, G., Li, T., Gao, J., 2011. A common parallel computing framework for modeling hydrological processes of river basins. Parallel Computing, 37 (6-7) 302-315.

Wang, X., and A. M. Melesse. 2005. Evaluation of the SWAT model's snowmelt hydrology in a northwestern Minnesota watershed. Trans. ASAE 48(4): 1359-1376.

Wiener, J., Gilmour, C. and Krabbenhoft, D., 2003. Mercury strategy for the bay-delta ecosystem: a unifying framework for science, adaptive management, and ecological restoration. Final Report to the California Bay Delta Authority. On the Internet at: http://www.science.calwater.ca.gov/pdf/MercuryStrategyFinalReport.pdf

Willems, P., 2006. Random number generator or sewer water quality model? Water Sci. Technol., 54(6-7) 387-394.

Willmott, C. J., 1981. On the validation of models. Physical Geography 2: 184-194.

Wood, M.L., Cooke, J., and Foe, C., 2006. Sacramento – San Joaquin Delta Estuary TMDL for Methylmercury, Staff Report. Available at: https://www.waterboards.ca.gov/centralvalley/water_issues/tmdl/central_valley_projects/delta_hg/archived_delta_hg_info/staff_report_jun06/

Yang, J., 2011, Convergence and uncertainty analyses in Monte-Carlo based sensitivity analysis. Environmental Modelling and Software, 26 (4), 444-457.

Zhang, X., Srinivasan, R., Arnold, J., Izaurralde, R.C. and Bosch, D., 2011. Simultaneous calibration of surface flow and baseflow simulations: a revisit of the SWAT model calibration framework. Hydrological Processes, 25(14), pp.2313-2320.