6. EVALUATION PLAN FOR TRIM.FaTE

TRIM.FaTE is a predictive environmental fate and transport model designed to support decisions on programmatic policy and regulation for multimedia air pollutants. These decisions can have far reaching human health, environmental, and economic implications. It is important that an assessment of how well the model is expected to perform the tasks for which it was designed is incorporated within the model development process. In other words, the trustworthiness of models used to determine policy or to attest to public safety should be ascertained (Oreskes et al. 1994). This chapter describes the role of model evaluation in developing an assessment of model quality and acceptability in support of regulatory decisions. The chapter provides background on the evolution of model validation terminology and concepts as well as previous Agency efforts (Section 6.1). The chapter then provides an introduction to model evaluation (Section 6.2) and presents an evaluation plan for TRIM.FaTE using four basic components (Sections 6.3 through 6.6). Finally, the Agency's progress in implementing the plan to date is described (Section 6.7).

6.1 BACKGROUND

Most of the early efforts to establish the quality of models used in supporting policy decisions focused on model validation. The term *validation* does not necessarily denote an establishment of truth, but rather the establishment of legitimacy (Oreskes et al. 1994). However, common practice has been not consistent with this restricted sense of the term, and the term validation has been commonly used in at least two ways: (1) to indicate that model predictions are consistent with observational data, and (2) to indicate that the model is an accurate representation of physical reality (Konikow and Bredehoeft 1992). The ideal of achieving – or even approximating – truth in predicting the behavior of natural systems is unattainable (Beck et al. 1997). As a result, the scientific community no longer accepts that models can be validated using ASTM standard E 978-84 (*i.e.*, comparison of model results with numerical data independently derived from experience or observation of the environment) and, therefore, be considered to be "true" (U.S. EPA 1998g). It is unreasonable to equate model validity with its ability to correctly predict the future (unknowable) true behavior of the system. A judgment about the validity of a model is a judgment on whether the model can perform its designated task reliably (*i.e.*, minimize the risk of an undesirable outcome (Beck et al. 1997)).

The current approach used by the Agency is to replace model *validation*, as though it were an endpoint that a model could achieve, with model *evaluation*, a process that examines each of the different elements of theory, mathematical construction, software construction, calibration, and testing with data (U.S. EPA 1998g). Therefore, the term *evaluation* will be used throughout this report to describe the broad range of review, analysis, and testing activities designed to examine and build consensus about a model's performance.

Over the last 10 years, the Agency has been considering model acceptance or model use acceptability criteria for selection of environmental models for regulatory activities. The Agency's efforts in this area are a result of SAB recommendations in 1989 that "EPA establish a general model validation protocol and provide sufficient resources to test and confirm models with appropriate field and laboratory data" and that "an Agency-wide task group to assess and guide model use by EPA should be formed" (U.S. EPA 1989). In response, EPA formed the Agency Task Force on Environmental Regulatory Modeling (ATFERM). This cross-agency task force was charged to make "a recommendation to the Agency on specific actions that should be taken to satisfy the needs for improvement in the way that models are developed and used in policy and regulatory assessment and decision-making" (Habicht 1992). In its March 1994 report, ATFERM recommended the development of "a comprehensive set of criteria for model selection (that) could reduce inconsistency in model selection and ease the burden on the regions and states applying the models in their programs," and they drafted a set of "model use acceptability criteria" (U.S. EPA 1994a).

More recently, an Agency white paper work group was formed to re-evaluate the recommendations in the 1994 ATFERM report. As a result, in 1998, EPA drafted the *White Paper on the Nature and Scope of Issues on Adoption of Model Use Acceptability Guidance* (U.S. EPA 1998g), which recommends the use of updated general guidelines on model acceptance criteria (to maintain consistency across the Agency) and the incorporation of the criteria into an Agency-wide strategy for model evaluation that can accommodate differences between model types and their uses. The work group also recommended the initial use of a protocol developed by the Agency's Risk Assessment Forum to provide a consistent basis for evaluation of a model's ability to perform its designated task reliably. The *White Paper* was reviewed by SAB in February 1999, and it is currently being revised in respond to SAB comments. The proposed approach for evaluation of TRIM.FaTE, as described in the evaluation plan presented here, is intended to be consistent with the Agency's current thinking on approaches for gaining model acceptability.

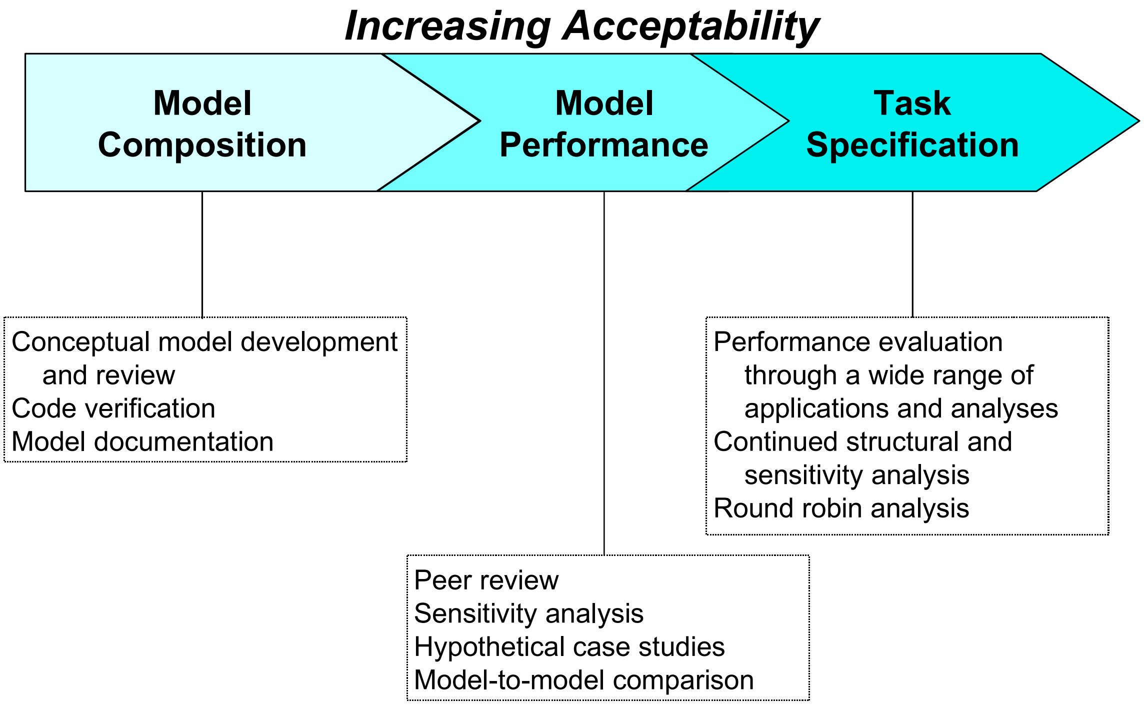
In its May 1998 review of TRIM.FaTE, SAB recognized the challenge in developing a methodological framework for evaluating a model such as TRIM.FaTE. Further, SAB suggested that "novel methodologies may become available for quantitatively assuring the quality of models as tools for fulfilling specified predictive tasks" (U.S. EPA 1998a). In developing the evaluation plan for TRIM.FaTE, the Agency has attempted to incorporate the essential ingredients for judging the acceptability of TRIM.FaTE for its intended uses, while retaining the flexibility to accommodate new methods that become available or changes in direction indicated by knowledge gained through the evaluation process.

6.2 MODEL EVALUATION

Model evaluation is necessary to increase the acceptance of a model. It is not a one-time exercise but a continuing and critical part of model development and application. Several model evaluation methods have emerged in recent years (Dennis et al. 1990, Hodges and Dewar 1992, U.S. EPA 1994b, Cohn and Dennis 1994, Eisenberg et al. 1995, Spear 1997, Schatzmann et al. 1997, Beck and Chen 1999, Arnold et al. 1998, Chen and Beck 1998). All of these methods can be placed into one of two basic categories: (1) those that focus on the performance or output from the model, and (2) those that test the internal consistency (Beck et al. 1997, Beck and Chen 1999) or scientific credibility (Eisenberg et al. 1995) of the model. These methods range from objectively matching model output with monitoring data to more subjective and abstract quality measures (*e.g.*, expert judgment, peer review). The focus of model evaluation activities will change during the life of a model. As a model matures, less emphasis will be placed on peer review and internal consistency checks and more resources will be directed toward evaluating how well the model satisfies both its original design objective and the specific modeling objectives of individual users.

Model evaluation can be viewed as a consensus building process (Figure 6-1) including three aspects as identified by Beck et al. (1997) – (1) model composition, (2) model performance, and (3) task specification – and recognized in the Agency's December 1998 *White Paper* (U.S. EPA 1998g).

Figure 6-1
Conceptual Representation of the Model Evaluation Process


Increasing Acceptability
Model Composition
Model Performance
Task Specification
Conceptual model development and review
Code verification
Model documentation
Performance evaluation through a wide range of applications and analyses
Continued structural and sensitivity analysis
Round robin analysis
Peer review
Sensitivity analysis
Hypothetical case studies
Model-to-model comparison

The evaluation plan for TRIM.FaTE is presented in the following four sections of this chapter, which correspond to different (but overlapping) types of model evaluation activities:

- Conceptual model evaluation;
- Mechanistic and data quality evaluation;
- Structural evaluation; and
- Performance evaluation.

The first three primarily focus on the information that goes into the model (*e.g.*, theory and data); how this information is synthesized (*e.g.*, process models, assumptions, and algorithms); and how the finished model is set up (*e.g.*, relevant level of complexity). The fourth focuses mainly on the information that comes out of the model (*e.g.*, comparing overall model outputs to various kinds of benchmarks).

The model evaluation plan designed for TRIM.FaTE must be flexible. Results from initial TRIM.FaTE evaluation efforts are posing new questions and leading to additional review, analysis, and testing. The various evaluation activities performed on TRIM.FaTE increase the experience and understanding that will ultimately lead to a judgment about its quality, reliability, relevance, and acceptability. The activities that are currently part of the consensus building process for TRIM.FaTE are described in the following sections. A number of these activities have been completed or are underway (*e.g.*, code verification, model documentation, peer review, case studies, sensitivity analysis), while others are still in the conceptual or planning stages.

EVALUATION THROUGHOUT MODEL DEVELOPMENT

As noted in the text, model evaluation is being performed in conjunction with model development. The evaluation activities performed to date have used the most current Prototype (*i.e.*, I through V) of TRIM.FaTE available at the time. Activities since the May 1998 SAB meeting have focused on Prototype V. These evaluation activities are fully applicable to TRIM.FaTE Version 1.0, which is being built from the same simulation algorithms and data as Prototype V. After verification that Version 1.0 produces identical results to Prototype V, Version 1.0 will become the focus of future model evaluation activities.

6.3 CONCEPTUAL MODEL EVALUATION

6.3.1 DEFINITION AND GENERAL APPROACH

Conceptual model evaluation is initiated in the early stages of model development. During the process of framing the problem and designing the conceptual model, the appropriate level of modeling complexity (*e.g.*, what to include and what to exclude), the availability and quality of information that will be used to run the model (*i.e.*, input data), and the theoretical basis for the model should be evaluated. A literature review should be undertaken to identify and evaluate the state-of-the-science for processes to be included in the model, as well as to compile and document the initial set of values that will be used as model inputs.

Conceptual model evaluation activities focus on the theory and assumptions underlying the model. These activities seek to determine if the model is conceptually sound.

Examples of conceptual model evaluation activities include:

- Literature review;
- Model documentation; and
- Peer review of problem definition and modeling concepts and approaches.

6.3.2 TRIM.FaTE-SPECIFIC ACTIVITIES

Considerable progress has been made in developing, documenting, evaluating, and refining TRIM.FaTE, including the following.

- An initial literature review identifying the state-of-the-science and the rationale for development of TRIM.FaTE has been completed (U.S. EPA 1997b, U.S. EPA 1997c), and the problem and design objective have been clearly defined (U.S. EPA 1998e).

- Model documentation has been extensive. TRIM Status Reports have been published in 1998 (US EPA 1998e) and 1999 (this document), and presentations have been made at scientific meetings including the Society of Environmental Toxicology and Chemistry (SETAC) annual meetings in 1997 (McKone et al. 1997a, Zimmer et al. 1997, Efroymson et al. 1997) and 1998 (Vasu et al. 1998) and the Society for Risk Analysis (SRA) in 1997 (Vasu et al. 1997, Guha et al. 1997, Lyon et al. 1997, Bennett et al. 1997, McKone et al. 1997b, Johnson et al. 1997). A detailed Technical Support Document for TRIM.FaTE is available (U.S. EPA 1999i, U.S. EPA 1999j), updated from a previous version (U.S. EPA 1998f).

- A May 1998 review by the SAB has been published (U.S. EPA 1998a). Additional evaluations of the conceptual model will continue to be reported in peer reviewed journals and will be subject to additional SAB consultation and review.

As refinements to TRIM.FaTE are made and as new applications are performed, conceptual model evaluation will continue.

6.4 MECHANISTIC AND DATA QUALITY EVALUATION

6.4.1 DEFINITION AND GENERAL APPROACH

Mechanistic and data quality evaluation activities focus on the specific algorithms and assumptions used in the model. These activities seek to determine if the individual process models and input data used in the model are scientifically sound, and if they properly "fit together."

Multimedia fate models are built around a series of process models (*i.e.*, algorithms or groups of algorithms) that make up the mechanics of the model. Many individual process models are taken directly from the literature and have been tested previously for performance and peer reviewed. The prior testing and review provides a degree of confidence that the process model correctly captures the behavior of the processes it is intended to model. New process models and assumptions are often introduced during model development; these new components need to be evaluated individually to ensure that they are working properly.

Mechanistic and data quality evaluations help to elucidate the internal workings of the model and, when necessary, provide a basis to refine process models and assumptions that play a critical role in the calculations. Sensitivity analysis methods are used to identify important model inputs during mechanistic evaluations and to identify the process models having the greatest influence on the model output. For example, alternative algorithms for the same process can be modeled and the results compared. Similarly, each time the model is used for a new kind of application, a sensitivity analysis may be appropriate to identify inputs, algorithms, and assumptions that have the greatest influence on the model outcome in that application. The quality and reliability of these influential factors directly affect the quality and reliability of the outcome from the analysis (Maddalena et al. 1999, Taylor 1993). When feasible, these influential factors should be refined to provide the best inputs to the analysis or, at the very least, identified as a potential source of uncertainty in the outcome.

Some mechanistic and data quality evaluation activities consider the model in its entirety. Process models are typically developed and tested in controlled or simplified systems. Therefore, how well these individual process models will perform in a fully coupled system is unknown. Mechanistic and data quality evaluations are designed and used to measure certain bounded indices of performance (*e.g.*, mass balance, appropriate and realistic mass transfer rates, relative concentrations within reasonable bounds). In addition, algorithms or routines that are used in a model to manipulate the data or to solve a system of equations (*e.g.*, LSODE, the differential equation solver used in TRIM.FaTE) need to be tested during the mechanistic evaluation to ensure proper performance.

Examples of mechanistic and data quality evaluation activities include:

- Computer code verification;
- Verification of generic algorithms adapted for and used within a model;
- Literature review to determine the extent of prior process model testing;
- Peer review of model components;
- Sensitivity analysis to identify important process models;
- Mass or molar balance checks;
- Performance evaluation of new and existing individual process models and of multiple process models in a linked system (*e.g.*, compare with existing models or with measurements, when available);
- Comparison of alternative process models (*e.g.*, equilibrium versus bioenergetic model for fish bioaccumulation of mercury);
- Data acquisition and evaluation (*e.g.*, data quality or reliability relative to the other inputs and assumptions), and development and documentation of default input data;

- Distribution development for input data to support probabilistic analysis; and

- Generic sensitivity analysis to help identify parameters that are most influential on model results, as well as potential data limitations (*i.e.*, model inputs that need further refinement or that are potential sources of uncertainty in the analysis).

6.4.2 TRIM.FaTE-SPECIFIC ACTIVITIES

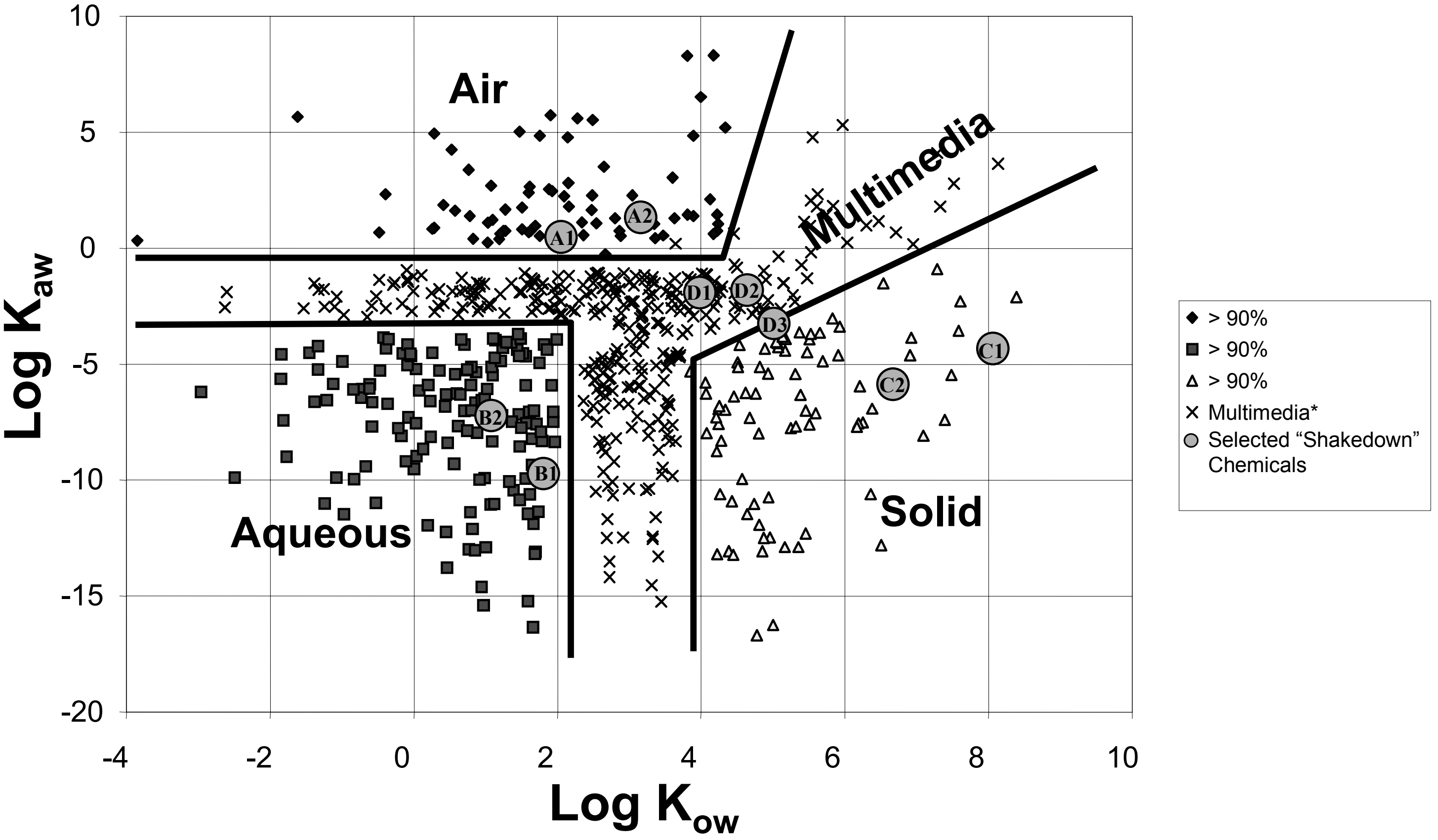
Prototype V (*i.e.*, spreadsheet-based model) is the current working version of TRIM.FaTE, and Version 1.0 (*i.e.*, Java-based platform) is under development (see Chapter 10). One of the features of TRIM.FaTE Prototype V that aids in mechanistic and data quality evaluation is its web-based output functions. There is an option to create a "full-recursive output," which documents the mass flow, as well as the associated transfer factors, to and from each compartment. The equation for each transfer factor can be viewed on a separate web page, and any calculated quantities used in that equation can then be viewed on additional pages. In this manner, checks can be made to ensure that the equations are input properly, and that the computer code is correctly calculating intermediate values. Analyses have been conducted on various parts of the code using this function.

In addition to the standard computer code verification efforts, performance of the generic code used to solve the differential equations in TRIM.FaTE (*i.e.*, LSODE) has been reviewed. Mass and molar balance checks are incorporated in the model for non-transforming organic compounds and mercury to allow for the quick assessment of model performance under a range of conditions.

Prior to conducting detailed evaluations of TRIM.FaTE's process models, numerous model runs were performed. It was determined that there was too much information in a complete run to evaluate whether the model was producing results that are logical, internally consistent, and reasonable. Thus, a "shakedown" phase of the model evaluation was begun using a set of hypothetical chemicals with a broad range of chemical properties. These hypothetical chemicals were designed to systematically probe the model across the broadest range of fate scenarios. The environment in its simplest form can divided into three major phases (*i.e.*, solid, aqueous, and gaseous). The relative solubility of a chemical in each of these phases indicates much about where and how the chemical will partition when released to the environment. These three solubilities can be represented by two fundamental partition coefficients, K_{ow} (*i.e.*, octanol/water partition coefficient) and K_{aw} (*i.e.*, non-dimensionalized Henry's Law constant, air/water partition coefficient).

A simple, level three (steady-state) mass balance model was used to identify the environmental phases for a randomly generated set of 500 "pseudochemicals" plotted in Figure 6-2. From this plot, the regions of parameter space that result in predominantly (>90 percent) single medium pollutants or multimedia pollutants can be identified. Two chemicals from each

Figure 6-2
Single Medium and Multimedia Chemical Regions for 500 "Pseudochemicals"



* Multimedia defined as not more than 80% of total mass in any single medium. Chemicals between 80% and 90% of total mass in any single medium were excluded from selection as "shakedown" chemicals.

single medium pollutant class and three from the multimedia pollutant class were selected for use as the initial shakedown evaluation set. This approach is particularly useful when performing diagnostic evaluations because the set of pseudochemicals provides insight into possible reasons for unexpected model outcomes. For example, if the model predicts an unusually high concentration in plants for the gas phase chemicals while the aqueous, solid, and multimedia phase chemicals seem reasonable, a problem in one of the diffusion algorithms would be suspected. Often, the model is run with only a subset of the available compartment types to focus on a particular algorithm or set of algorithms. To date, this group of shakedown chemicals has been used to evaluate and debug the soil algorithms, the plant algorithms, and the general biotic algorithms. These pseudochemicals will continue to be used to further evaluate the process models in TRIM.FaTE and the model as a whole.

Tests are being performed or designed to evaluate process models that, according to the literature review, have not been thoroughly tested, as well as for approaches and algorithms developed specifically for TRIM.FaTE. Examples of process models that have been identified for evaluation include the particle/plant leaf algorithms, the soil flux model, and the air transport algorithms. Approaches and algorithms that are related to seasonality (*e.g.*, snow, plant growth, senescence) will be evaluated so that they can be incorporated into TRIM.FaTE, if appropriate.

When different models are available for the same process (*e.g.*, bioaccumulation in fish), model-to-model evaluations may be performed at a process model level to test the overall performance of TRIM.FaTE using different input algorithms. As one example of this, EPA is comparing the air transport component of TRIM.FaTE to a widely used EPA air dispersion model, ISCST3 (U.S. EPA 1995c). In addition, measured concentrations that are available for a single medium or multiple adjacent media (*e.g.*, water and sediment, or water and fish) will be used, where available, to test single or multiple process models.

Data acquisition and the careful evaluation of model inputs are ongoing. To date, the majority of effort has focused on compiling an initial set of model inputs for a small set of test chemicals (*i.e.*, phenanthrene, benzo(a)pyrene, mercury) and environmental settings (U.S. EPA 1998f; also Chapter 7 and Appendix C of this document).

In addition, sensitivity analysis techniques are being used to provide a first-order determination of the most influential parameters in TRIM.FaTE. The sensitivity of model outputs to changes in individual parameters is assessed by performing a series of simulations where each parameter is varied with the other parameters held constant. This does not take into account parameter dependencies or synergistic effects, but is an efficient way to perform an initial assessment of the relative influence of the parameters. This information supports model evaluation by providing a prioritized list of parameters on which to focus the evaluation efforts.

To take full advantage of the probabilistic capabilities of TRIM, some inputs will need to be described using probability distributions that separate uncertainty and variability. The uncertainty and variability analysis methodology that has been developed for TRIM.FaTE is further described in Section 4.7 and in TRIM.FaTE TSD Volume I, Chapter 6. Following the implementation of this methodology, sensitivity analyses are being performed to help identify potential influential factors and data limitations. One of the key functions of the uncertainty analysis methodology is to evaluate the importance, in terms of both uncertainty and variability, of specific model inputs and of model components in relation to other inputs and components. This gives insight into priorities for reducing uncertainty and for focusing efforts on the improved representation of model inputs. The ability to rank input parameters in order of their influence on the uncertainty and variability of the model results is an important component of establishing such priorities.

As refinements to TRIM.FaTE are made and as new applications are performed, data quality evaluation will continue to be revisited. Sensitivity analysis can be used to identify inputs, algorithms, and assumptions that have the greatest influence on the model outcome in specific applications. When feasible, influential factors may be refined to provide the best inputs to the analysis or identified as a potential source of uncertainty in the outcome.

6.5 STRUCTURAL EVALUATION

6.5.1 DEFINITION AND GENERAL APPROACH

Judging the reliability of a model requires an understanding of how the model responds to changes in complexity (*i.e.*, changes in the modeling structure). Both temporal and spatial changes can be made to the model structure. Structural evaluation addresses these kinds of changes and provides valuable information about the performance and behavior of the model under a range of conditions, improving the ability to judge the model's quality and reliability. Ideally, these evaluations can help determine the optimal model structure to balance the level of complexity needed to create reliable outputs with the simplifications that can make the model easier and more practical to use. If the model is less complex, it is easier to perform additional analysis, such as uncertainty and sensitivity analysis, and is more practical to apply to specific sites and situations. Structural evaluation can provide insight and guidance for future model applications, and it is a very useful input to developing user guidance.

Structural evaluation activities focus on how changes in modeling complexity affect model performance. These activities seek to determine how the model will respond to being set up differently for different applications.

A large number of well designed runs is necessary to evaluate the way a model performs under different conditions. These structural evaluations combine sensitivity analysis methodology with model-to-model comparisons. For a structural evaluation, the model is set up for an application, using either real or hypothetical data. Changes are then made to the structure (*e.g.*, spatial elements are split or recombined; time steps are changed; compartment shapes, sizes, and locations are altered), and the model outcomes are compared (*i.e.*, the model is compared to itself under various set-up conditions).

Structural evaluations encompass a series of comparisons designed to measure the model's response to various changes, which can include:

- Different run duration and/or time step values;

- Varying spatial configurations;

- Changes in initial and boundary chemical concentrations;

- Changes in the source and/or target locations; and

- Other changes in complexity (*e.g.*, including/excluding biota, using average precipitation versus discrete rain events).

6.5.2 TRIM.FaTE-SPECIFIC ACTIVITIES

TRIM.FaTE is intended to be used in a wide range of modeling applications (*e.g.*, various chemicals, environmental settings, exposure conditions). Because TRIM.FaTE can be used at various levels of complexity, it is important to understand the level of complexity needed for a particular analysis and the stability of model output when the system structure is changed. Given the complexity of the "real world" and the large number of inputs used in TRIM.FaTE, a complete set of structural evaluations cannot be identified and performed. The focus of structural evaluation activities for TRIM.FaTE will be responsiveness to changes in model complexity with respect to both temporal and spatial scales and the types of compartments included.

Several structural evaluation activities have been identified for TRIM.FaTE, including the following.

- **Response of abiotic compartments to the exclusion/inclusion of biota.** It has typically been assumed that the mass of a chemical in biota compared to the mass in abiotic compartments (*e.g.*, soil, water, air) is not large enough to influence the overall chemical mass balance. However, if both the flux into the biotic compartment and the reaction rates within the compartment are rapid enough, the biota can potentially influence the mass balance even when a relatively small volume of biota is present (Maddalena 1998). Testing will be done to measure the model response to biota inclusion to determine when and to what extent biota need to be included in mass balance calculations.

- **Response to temporal scales of analysis and to aggregate inputs.** Detailed meteorological data are available and will be used in a simplified scenario, as part of the mercury case study (see Chapter 7), to test the model's response to aggregation of input data over various time periods. By running the model with varying degrees of input aggregation, the level of input detail required to achieve a specified level of detail in the output can be determined.

- **Response to changes in the size, shape, and location of compartments.** As part of the mercury case study (see Chapter 7), EPA plans to examine the effect of varying spatial configurations on TRIM.FaTE results. This will include changing the size of compartments in multiple dimensions to determine the most appropriate way to grid the model, as well as adding compartments at the edges of the model region to examine the boundary effects around the model system (*i.e.*, flux of chemical mass into or out of the system).

Results from initial structural evaluation analyses would likely lead to further testing (*i.e.*, diagnostic evaluations). Different tests could be designed and executed until a clear understanding of the behavior of TRIM.FaTE at different levels of complexity emerges. This understanding will ultimately be incorporated into a user's manual to provide guidance on setting up the model at an appropriate level of complexity for a given application. For practical reasons, it is important to limit the complexity of model setup to that which is needed to produce acceptable modeling results.

6.6 OVERALL PERFORMANCE EVALUATION

6.6.1 DEFINITION AND GENERAL APPROACH

Performance evaluation activities focus on the output of the full model. These activities seek to determine if the output is relevant, reliable, and useful.

Unlike the other types of model evaluation discussed above, which focus on specific aspects of the model (*e.g.*, inputs, process models), performance evaluation focuses on the model as a whole. Performance evaluation compares modeling results to some type of benchmark (*e.g.*, monitoring data, other modeling results). Generally, various performance evaluation analyses are conducted in a similar manner, with only the source of the comparison data changing. The optimized model, as modified based on all prior evaluations, is used for performance evaluation.

Matching model output to monitoring data is often considered the most desirable form of performance evaluation. Although comparing model output to measured values provides useful information on the model, history matching experiments provide only part of the overall picture of the model's quality, reliability, and relevance (Beck et al. 1997). Several other forms of performance evaluation exist. In addition to monitoring data, output of another model and expert opinion and judgment about how output should look can be used as comparison benchmarks in performance evaluation.

Moreover, each evaluation provides an opportunity to use the model. In addition to the ultimate findings of the performance evaluation itself, the experience gained through these exercises contributes to an overall understanding of the model, which ultimately enables both model developers and users to judge the quality of the model.

A different form of performance evaluation is the "round-robin" experiment (Cowan et al. 1995), in which several different users independently set up the model and generate output using the same data for a particular case study (*e.g.*, site description, chemical properties). Model outputs are then compared, and the users' experiences are reviewed to identify weaknesses and ambiguities in the program's user interface and other user guidance that could lead to errors inapplying the model. The lessons learned can then be incorporated into user guidance to help prevent user errors and inappropriate model applications.

Examples of performance evaluation activities include:

- Comparison of model output to monitoring data (*e.g.*, concentrations in environmental media and biota, exposure markers);

- Model-to-model comparison;

- Round-robin experiments; and

- Some forms of regional sensitivity analysis (*i.e.*, output is tested against knowledge about a plausible bound).

6.6.2 TRIM.FaTE-SPECIFIC ACTIVITIES

An extensive review of the literature was undertaken following SAB's initial comments on the importance of model evaluation for the TRIM project (U.S. EPA 1998a). The review focused on identifying potential data sets for use in evaluating the performance of TRIM.FaTE. The usefulness of some of the reported environmental measurements was limited because in many cases the source of the chemical contamination was not well characterized. Several studies were identified that report chemical measurements in multiple environmental media (Table 6-1). The majority of these studies focus on measuring the current chemical concentrations in the environment with little emphasis on temporal variability or trends. Several of the studies were designed to assess multimedia partitioning (*e.g.*, atmospheric partitioning among the gas, aerosol, and water phases) or to investigate specific environmental processes such as the transfer rate across an environmental interface. Although historical emission patterns can potentially be reconstructed for certain chemicals using sediment chronology (Cowan et al. 1995), little effort has gone into matching historical emissions to multimedia environmental concentrations.

None of the studies identified during EPA's literature review provides complete and concurrent information on chemical concentrations in the five major environmental media (*i.e.*, air, water, sediment, soil, biota) along with the associated source term(s) and environmental characteristics (*e.g.*, meteorology, hydrology, landscape properties). Although some of these studies can and will be used to evaluate certain aspects of the model, it is important not to overvalue these results when judging the overall quality of the model.

As noted above, comparisons of TRIM.FaTE outputs to monitoring data are difficult because complete multimedia data sets from well-characterized systems (*e.g.*, known source, meteorology, and landscape) to use in a performance evaluation are not currently available. However, limited data sets are becoming available through the literature (see Table 6-1) and through unpublished sources (*e.g.*, multimedia monitoring by state or local agencies). These smaller data sets will allow TRIM.FaTE's output to be evaluated and compared with measurements, at least to some degree.

Table 6-1
Multiple Environmental Media Studies

Chemical	Speciation?	Source	Location	Media Measured	Sampling Frequency	Study
Benzo(a)pyrene, other PAHs	NA	Urban	Florence, Italy	• Air particulate • Plant	Once	Ignesti et al. (1992)
Benzo(a)pyrene, PAHs (4), PCBs	NA	Petrochemical factories	Stenungsund, Sweden	• Plant • Soil	Once	Thomas et al. (1984)
Chlorpyrifos	NA	Not specified	Chesapeake Bay	• Air • Water	1993 (four times per year from eight stations)	McConnell et al. (1997)
Dioxins	NA	Urban	Bolsover, Derbyshire, England	• Air (including deposition rate) • Plant	Once	Jones and Duarte-Davidson (1997), Duarte-Davidson et al. (1997)
Mercury	In mammals and earthworms only	Chloralkali plant	Great Britain	• Air • Earthworm • Grass • Soil • Wood mouse and vole organs	Once	Bull et al. (1977)
	None	Lithium separation facility	Oak Ridge, TN	• Earthworm • Grass • Mouse • Shrew • Soil	Once	Talmage and Walton (1993)
	None	Chloralkali plant	India	• Goat • Some plant species parts • Sheep • Soil	Once	Shaw and Panigrahi (1986)
	None	Chloralkali plant	India	• Aquatic plant • Crop plant • Soil • Sediment • Water	Once	Lenka et al. (1992)

Chemical	Speciation?	Source	Location	Media Measured	Sampling Frequency	Study
Mercury (continued)	None	Chloralkali plant	Italy	• Air • Soil • Plant	Once	Maserti and Ferrara (1991)
	None	Cinnabar, mining	Italy	• Air • Rain water • Plant • Soil • Surface water sediment	Once	Ferrara et al. (1991)
	Some methylmercury	Chloralkali plant	Saltmarsh ecosystem near Brunswick, GA	• Birds • Fish • Invertebrates • Mammals • Plant Parts • Sediment	Once	Gardner et al. (1978)
	Total, methyl, dissolved gaseous	Urban/runoff	Chesapeake Bay and streams	• Precipitation • Sediment • Water	Several single event measurements (1995 through 1997)	Mason et al. (1999, 1997a,b)
Metals, pesticides, PAHs	NA	Not specified	Two different regions in US	• Air (indoor and outdoor) • Biologic fluid • Food (market basket) • Soil	Single measurements per household (early 1990s)	U.S. EPA (1999a), Sexton et al. (1995)
	NA	Not specified	Northeastern US	• Air (indoor and outdoor) • Biologic fluid • Food (market basket) • Soil	Longitudinal study of several households (early 1990s)	U.S. EPA (1999a), Sexton et al. (1995)
MTBE	NA	Multiple estimated	California	• Air • Ground water • Surface water	1997-98 and prior	University of California (1998)
Organochlorines	NA	Not specified	Lake Baikal, Russia	• Fish • Seal • Water (dissolved and particulate) • Zooplankton	1993 (August - September)	Kucklick et al. (1996)

Chemical	Speciation?	Source	Location	Media Measured	Sampling Frequency	Study
Organochlorines (continued)	NA	Not specified	Lake Superior	• Aquatic biota (19, from amphipod to lake trout)	Summer 1994 (at multiple sites)	Kucklick and Baker (1998)
	NA	Urban	Lake Michigan	• Precipitation	Summer 1994 (at multiple sites)	Offenberg and Baker (1997)
Organochlorines, PAHs	NA	Not specified	Chesapeake Bay and streams	• Air (vapor and aerosol) • Atmospheric deposition • Diffusive exchange • Water (dissolved and suspended particles) • Plankton • Wet deposition	October 1990 - August 1992 (at multiple sites over, in, and adjacent to lake)	Ko and Baker (1995), Leister and Baker (1994)
PAHs (10)	NA	Urban	Indiana	• Air (particulate) • Gas • Plant	Every 20-30 days for several months	Simonich and Hites (1994)
Total PAH	NA	Road	Australia	• Air (particulate) • Grass • Soil	Once	Yang et al. (1991)

An important aspect of the plan for performance evaluation of TRIM.FaTE is a detailed case study of a mercury-emitting industrial facility, which was chosen in part because of the availability of multimedia monitoring data and concurrent emission estimates from a local source. The mercury case study site also is playing a critical role in the mechanistic and data quality, and structural evaluations being done, as well as serving as the basis for a variety of sensitivity analyses. Chapter 7 describes the mercury case study, including the available environmental and biotic measurement data, in more detail.

The previous prototype of TRIM.FaTE was compared with two similar models, CalTOX (McKone 1993a, McKone 1993b, McKone 1993c) and SimpleBox (van de Meent 1993, Brandes et al. 1997). The pollutants modeled for this comparison were PAHs (U.S. EPA 1998f). More recently, outputs from TRIM.FaTE are being compared to outputs from the EPA's ISCST3 and IEM2M models, as part of the mercury test case (see Chapter 7). ISCST3 will be used to generate air deposition and concentration data that will be used in IEM2M to estimate multimedia concentrations of mercury. These concentrations will be compared to TRIM.FaTE outputs that will be modeled using consistent inputs, as well as to TRIM.FaTE outputs from an analysis where the air depositions and concentrations from ISCST3 are incorporated into TRIM.FaTE (in essence, substituting for TRIM.FaTE's air transport component). As part of the mercury test case, TRIM.FaTE outputs (*e.g.*, ranges of predicted environmental media and biotic concentrations of mercury) will also be compared to the available measurement data for mercury in environmental media and biota. The predicted ranges of model results used for these comparisons will be based on the results of TRIM.FaTE uncertainty and variability analyses, as described in Chapter 6 of TRIM.FaTE TSD Volume I.

Although most model-to-model comparisons are performed on a scenario-specific basis, a more informative approach may be to compare models across a range of conditions using multiple regression or data mining software (Helton et al. 1989, Spear et al. 1994). In the future, more robust forms of model-to-model comparison may be considered for TRIM.FaTE.

Sensitivity analyses are often used in performance evaluations to identify the part of the model that is actually being tested. Given the large number of inputs used in multimedia models such as TRIM.FaTE, it is not always obvious which processes and algorithms are participating in the calculation. TRIM features for uncertainty and variability analysis (see Chapter 3), standard sensitivity analysis methods, and regional sensitivity or parameter space analysis methods (Beck and Chen 1999, Spear 1997) may be used to understand and communicate the results from performance evaluations and to improve the ability to assimilate the results from all the evaluation efforts.

6.7 SUMMARY OF TRIM.FaTE EVALUATION ACTIVITIES

The TRIM.FaTE evaluation plan, as described in this chapter, includes a variety of activities designed to build consensus about the model's performance and increase acceptance of the model for its intended applications. A few of these activities have been completed, many are in progress, and several others are in the planning stages. Table 6-2 summarizes key elements of the evaluation plan by providing examples of TRIM.FaTE evaluation activities to date as well as examples of planned future activities.

Table 6-2
Summary of TRIM.FaTE Evaluation Activities

Type of Evaluation	Evaluation Activity	Examples of Progress to Date	Examples of Future Plans
Conceptual Model Evaluation	Literature review	Extensive during model conceptualization and early development	Perform targeted reviews when adding or refining algorithms
	Model documentation	Status Reports and comprehensive TSDs in 1998 and 1999, presentations at scientific meetings	Update and expand documentation throughout development; develop user guidance
	Peer review of modeling concepts and approaches	Reviewed by SAB in 1998; full internal EPA review and SAB advisory in 1999	Periodic internal and external peer review
Mechanistic and Data Quality Evaluation	Computer code verification	Extensive during development for Prototypes I to V and Version 1.0; performed review of LSODE; compared Prototype V and Version 1.0 results for some test cases; developed automated tests of internal functions for Version 1.0	Complete comparisons between Prototype V and Version 1.0 results and reconcile any differences; develop and evaluate additional Version 1.0 internal tests
	Performance evaluation of process models that are components of TRIM.FaTE	Compared TRIM.FaTE to CalTOX output for nine "pseudochemicals" (*i.e.*, varying K_{ow}/K_{aw}) in a simple scenario (*i.e.*, air, water, soil); compared TRIM.FaTE to ISCST3 for air transport of mercury	Continue performance evaluation for process models (*e.g.*, particle/plant leaf algorithm, soil flux model)
	Comparison of alternative process models	Compared chemical flux across soil/air interface with results from Jury model; comparing chemical transfer from soil to root with physically based model	Compare K_{oa} (*i.e.*, octanol/air partition coefficient) aerosol model with the Junge model; perform model-to-model evaluations for bioaccumulation in fish models
	Data acquisition and evaluation/ development and documentation of default input data	Compiled an initial set of data for test chemicals (phenanthrene, benzo(a)pyrene, mercury) and environmental settings	Continue data acquisition and evaluation (*e.g.*, other chemicals and environmental settings)
	Generic sensitivity analysis of input parameters	Some analyses of Prototypes I to IV; initial analyses to determine elasticities of >100 parameters for Prototype V using mercury case study	Assess the most influential input parameters as part of future evaluations and applications

Type of Evaluation	Evaluation Activity	Examples of Progress to Date	Examples of Future Plans
Structural Evaluation	Analysis of time step resolution and other time-related aspects of modeling as part of case study	Very limited analysis	Perform detailed analyses; characterize variance due to temporal resolution changes in inputs; ensure that time-averaged output sufficiently maps the temporally resolved output
	Analysis of varying spatial configurations as part of case study	Limited analysis for air component only	Perform detailed analyses to characterize how robust the model is to spatial configuration changes
	Analysis of changes in complexity	Compared TRIM.FaTE for a simplified mercury case study scenario with and without biota	Identify issues to be addressed when setting up the model for an application
Overall Performance Evaluation	Regional sensitivity analysis	None to date	Identify regions of parameter space that are critical to certain model outcomes as part of future evaluations and applications
	Model-to-model comparison	Compared early prototypes to CalTOX and SimpleBox; have begun comparisons with ISCST3/IEM2M for mercury case study	Complete ISCST3/IEM2M for mercury case study comparisons
	Comparison to monitoring data	Have begun multimedia comparisons for mercury case study	Complete mercury case study; identify other test chemicals and sites, as needed

[This page intentionally left blank.]