Review of Model Validation: Modeling Study of PCB Contamination in the Housatonic River, Volumes 1 and 2

DCN: GE-030706-ADBR
March 2006

Background and Disclaimer

The review was conducted under a grant from the Environmental Protection Agency (EPA) to the Housatonic River Initiative. The technical and professional positions presented here are those of the author and do not represent the positions of the EPA, the U.S. Army Corps of Engineers, or any other federal or state agency.

This report was prepared by Steven W. Effler, Director of Research of the Upstate Freshwater Institute, Syracuse, NY, under contract to the Housatonic River Initiative.

Overview

I have reviewed both Volumes 1 and 2 of the document listed above for the Housatonic River. I have also reviewed my earlier comments made on the "model calibration" report in 2004; a copy is attached as Attachment No. 1.

I was supportive of much of the modeling approach and its execution in the calibration effort, as described in the earlier (2004) review, though a number of criticisms were also presented (see Attachment No. 1). Indeed, much of the validation effort demonstrates good modeling practices. However, the validation effort is seriously flawed by the lack of data to test model performance against. In particular, there is essentially no runoff event data, outside of the Phase 2 calibration data (or even Phase 1 calibration data), to test predictions of TSS and PCBs (e.g., t-PCBs). The following section expands upon this concern.

The Problem of the Lack of Runoff Event TSS and t-PCB Data for Validation

I expressed concern with the description of the supporting data set for the original calibration as "extensive" (see ¶ 1 at top of page 2, Attachment No. 1). Frankly, not only is the representation not valuable, it is inaccurate and misleading. As I indicated earlier … "Overall, I consider the data set **adequate** to support the presented modeling". That statement applied to the original 14 month calibration interval. The time expansion for Phase 2 calibration (1990 – June 2000) was reasonable to evaluate predictive capabilities for sediment issues. However, no substantive additional opportunities to test mobilization and transport of TSS and t-PCBs were embedded in the expanded period for Phase 2 testing.

The primary problem is the absence of runoff event observations of TSS and t-PCBs (i.e., multiple, closely-spaced in time, samples during events) within the validation interval, 1979 – 1990 and July 2000 through 2004. The lack of such data for the 1979 – 1990 interval is easily understood. But, why was event sampling for these constituents not done over the July 2000 through 2004 interval? The data may have been adequate for Phase 1 calibration, but data for the validation period are woefully inadequate.

This omission seems to be at fundamental "odds" with widely accepted modeling approaches. A representation of the components of the modeling process and the proper interactions between these components is presented as Attachment No. 2. The modeling process should effectively integrate modeling activities with monitoring and process studies, and allow for feedback/interaction (i.e., a dynamic process) between these components (see Attachment No. 2). Clearly the mobilization and transport of TSS and t-PCBs associated with runoff events is central to the issue of this contaminated site. Further, the systematic increases in these constituents during runoff events, and subsequent longitudinal patterns, offered some of the most conspicuous signatures for model calibration. These signatures represent opportunities to test a model. Ideally, model validation (e.g., verification in Attachment No. 2) should be conducted for substantially different conditions (e.g., drivers) with different signatures. In the absence of substantially different conditions (as is the case for this site), the model must be validated for more runoff events. This should have been obvious to the modeling team early in the modeling process; i.e., there was plenty of time to implement additional event sampling over the July 2000 through 2004 period. As a result of this omission, some of the primary signatures to test model performance are not available for this validation effort. The "validation" report seems to acknowledge the need for additional event sampling (p. 2-21, lines 13-20), listing "major storm event sampling" (line 19) as one of the five special studies that were either completed, in progress, or planned, in response to comments from the peer-review panel. What happened? If it's planned, will there be a Phase 2 validation?

Instead of an event-based sampling program for TSS and t-PCBs, a fixed-frequency program was conducted over the July 2000 through 2004 interval (e.g., Fig. 6.2-44). The rather poor model performance for t-PCB for validation (e.g. Fig. 6.2-58) is at least in part related to the failure to capture runoff events in sampling.

The modeling team's position that there are extensive data to support the modeling effort is not supported by the Validation Figures. Two hundred and sixty-six (266) pages of representations of model predictions are presented. Review of these pages leaves this reviewer with one over-riding impression – the small amount of data (particularly for the water column) available to support model validation; e.g., most pages with little or no data. This situation is at odds with the modeling "wisdom" shared in the report; e.g., that models are imperfect representations of reality. **Representative data are the reality**. Good and credible models are supported by rich and appropriate data sets. This site needs more data. Further, this professional (with more than 30 years experience in the field) is concerned with the "message" (even precedent) sent by this inadequacy for a very important clean-up site, on a broader geographic scale. Such important problems deserve more comprehensive scientific information.

Attachment No. 1

Review of "Model Calibration: Modeling Study of PCB Contamination in the Housatonic River"

DCN: GE-122304-ACMG
December 2004

by S. W. Effler

Review of "Model Calibration: Modeling Study of PCB Contamination in the Housatonic River"
DCN:GE-122304-ACMG
December 2004

Background and disclaimer

This review was conducted under a grant from the Environmental Protection Agency to the Housatonic River Initiative. The technical and professional positions presented here are those of the author and do not represent the positions of the EPA, ACOE or any other federal or state agency.

This report was prepared by Steven W. Effler, Director of Research of the Upstate Freshwater Institute, Syracuse, NY, under contract to the Housatonic River Initiative.

Overview

I have reviewed the entire **large** "model calibration" document listed above for the Housatonic River, and selected portions of certain of the attached appendices. Clearly the system/issues combination addressed is inherently complex, representing a challenge to available model frameworks. The selection of the three different frameworks, HSPF, and EFDC (with submodels), and FCM (previously approved in the overall site process), appears to be appropriate. Choices of time and space scales and segmentation (potential exception for EFDC, subsequently) for the individual models, and linkages between them, are appropriate. The setting of predetermined performance goals, while apparently an obvious good step is frankly a rare feature that has served here to nicely constrain the modeling process with respect to performance.

Overall I am supportive of the competency demonstrated in using these tools to address the difficult issues of this site. Modeling "savvy" is manifested for a number of different issues; good choices were made in most cases (some possible exceptions are addressed below). Model parsimony, the appropriate balance between complexity and credibility to meet the stated (and ambitious) goals, was, for the most part, well applied in the work. The report documents several of the overall "Modeling Study Objectives" being met; there is reason to believe the others will be met in subsequent modeling components. In general I support the position that targets for model performance have been met, without noteworthy bias in predictions, and that relevant processes have been appropriately accommodated. Further, reasonable sensitivity analyses have been conducted to identify the more important model inputs. Evidence of probabilistic treatment of model uncertainty was less even under the treatments of the various models. I assume this will become more unified as the modeling process continues. The modeling team made a genuine effort to minimize arbitrary "knob tuning" to achieve fits, though the extent of independent specification of inputs may have been somewhat "over-sold".

The description of the supporting data set as extensive is not particularly valuable and in fact open to debate. A large number of total observations is not really the point. A quick review of the various figures of the report results in numerous cases where it could be argued more data would be beneficial in evaluating performance. Conversely, there are cases (e.g., sediment PCB) where many more measurements than necessary were made. Overall, I consider the data set **adequate** to support the presented modeling.

I note that the longitudinal limits of this modeling analysis had been previously established, but that rather conspicuous impacts have been documented for the river through Connecticut. Is EPA prepared to extend the modeling effort through the entire impacted river reach. In particular, is an adequate monitoring program underway to support such an initiative.

Below in my specific comments I raise a number of other issues and questions that deserve consideration. These are presented in the order of appearance of material in the primary report. The most noteworthy concerns are highlighted with (*).

Specific Comments

A. Watershed Modeling (HSPF)

1. Development of model inputs was comprehensive and appropriate for flow (Q) simulations.
2. Attention to detail of spatial segmentation was appropriate.
3. Performance for Q simulations was very good for multiple time scales of annual, seasonal/monthly, base, and storm events (e.g., Tables 2-5, 2-6, 2-7; Figure 2-7 through 2-12); also applied to events outside of the May 99 through June 00 period.
4. Simulations of solids delivery by this, or any other widely available framework, is inherently much more uncertain; guidance of erosion calibration with USLE estimates is an appropriate approach.
5. Simulations of annual TSS loads are reasonably good (Table 2-12, Figure 2-12).

(*)6. TSS concentration vs. Q predictions (note log scales, (Figures 5-22 and 5-23) are only fair; better at New Lenox Rd (Figure 5-24).

7. Event TSS concentration simulation performance is fair/variable (Figure 2-25).
8. Performance targets were met, but these were much broader than Q because of acknowledged difficulties for TSS simulations.

(*)9. HSPF would not be a good choice for accurate stream temperature (T) simulation in most cases. However, given the modest goals for T model performance, and the relatively small effect of this source of uncertainty for the FCM model predictions, this application is supportable.

(*)10. The upstream T simulations (Figure 2-27) really are not good, while performance is substantially better downstream (Figure 2-28). Further, is there a disconnect here somewhere? Short-term

performance (e.g., Figure 2-29) for boundary simulations (e.g., Figure 2-29) seem systematically better.

(*)11. The simulated trapping efficiency for Woods Pond presented as the 3rd item at the bottom of p.2-57 - was this presented earlier in the text?

B. EFDC

1. Hydrodynamic Model

(*)a. General comments

I am aware EFDC is being adapted by EPA as a 3-D framework that EPA is to provide **some** support on. I have no problems with the choice; we are active users also, following a critical review of alternatives. An associated EPA user's manual still remains unavailable. Further, it should be reason for concern by EPA that most of the supporting references for the model continue to be in the "gray literature" category (e.g., p.3-40). This leaves this reviewer with the feel of "proprietary issues" embedded within what I understand to be a very "public" project. Perhaps this will sort itself out in the future.

(*)b. Model Grid.

Figure 3-2 does not present the EFDC grid. The grid has a very large number of cells. Given the very long model run time acknowledged later, what are the arguments for this fine spatial resolution. Is this fine resolution necessary to meet the overall model goals? Were any sensitivity analyses conducted - e.g., 0.5x, 0.25x this number of cells.

(*)c. The wetting/drying capability of EFDC makes it attractive for the floodplain issue. I am unclear on the vertical water level change necessary to make a floodplain segment active.

(*)d. Incorporation of macrophyte effects in downstream sections appears to be an innovative, and potentially important, feature, particularly with respect to sediment transport. I was disappointed not to see literature citations to support its development or previous incorporation in a model; nor do I recall a sensitivity analysis that demonstrates its effect on sediment transport in the river.

e. Model performance was generally good and within specified goals. The addition of simulation of spatial extent of out-of-bank coverage testing for Hurricane Bertha (Figures 3-12 and 3-13) and October 2003 (Figure 3-14) is a particularly valuable addition with respect to floodplain issues.

2. Sediment Transport Modeling

a. This submodel is complex. However, its structure appears to accommodate the processes identified in the MFD (bottom, p.4-1).

b. The use of "settling velocity" to lump the net effect of multiple complex processes for cohesive sediments is an appropriate simplification.

(*)c. Critical shear stress for resuspension was used as a calibration parameter (p. 4-5, line 19). Did this input vary in time and space? Guidance from the independent Sedflume experiments with model input values remains obscure.

d. Specification of sediment size classes is somewhat arbitrary.

(*)e. This model allows for a mixture of particle size classes within a cell. Should we be concerned that all the supporting citations are limited to the gray literature?

f. Upstream models were necessary to supply boundary conditions TSS according to multiple particle classes. Since particle size data were not routinely available, it relied on a single event (Oct. 03) for which partitioning was conducted.

g. The functionality presented in Figure 4-17 is primarily a manifestation of the operation of increasing flocculation/aggregation with increasing concentration. But if it represents the system it is fine.

h. Clearly there is substantial knob-tuning embedded in the calibration effort for the sediment sub-model. However, calibration testing was rigorous, particularly in the context of the various bases of performance evaluated (top of p.4-40).

(*)i. Most of the performance graphics present TSS concentration on log axes. I recommend some related statement be made … e.g., "consistent with common modeling presentations for this constituent".

(*)j. The performance in the time series format of Figure 4-25 (p.4-47) I would describe as only fair. Further, some bias is indicated, as the model overpredicted TSS in most cases.

k. Figure 4-28 is an excellent presentation, in that it allows a holistic view along the PSA of the imbalance between the erosion and deposition process for the four particle sizes classes. A reasonably consistent pattern emerges.

(*)l. Performance plots like the first and third panels of Figure 4-34 do not support the position of support by a rich data base. Further, performance is in some cases not impressive within the context of log-log plots. However, pre-established performances goals were met.

m. Appropriate sensitivity analyses were conducted. Further, these results were consistent with documented patterns and the gradients in physical character within the PSA.

(*)n. The ability to simulate longer-term (beyond stated 14 month calibration interval) erosion/deposition patterns, as reflected (in part) in the sediment record, could not be demonstrated over the short specified calibration interval. This is to be evaluated under the "validation" effort, through longer-term runs (p.4-69, line 16 and 17). This may require further model adjustment (i.e., calibration). Thus the "break" between the calibration and validation efforts remains imperfect. It's reasonable to ask what model simplifications will be made to facilitate long-term (multiple years) simulation, given the subsequently stated very long model run times.

3. PCB Fate and Transport Modeling

a. Figure 5-1 is effective in depicting the linkages and major component for the EFDC submodels and FCM.

(*)b. Frankly, "high-frequency" temporal and spatial data based on 11 events is

overstated.

(*)c. The "<" symbol on Figure 5-3, bottom panel, needs to be defined (detection limit value?).

(*)d. Given the uncertainty in bed load simulation (e.g., supported by direct measurements for one 2002 event), the associated PCB load must be considered also uncertain (p.5-8).

e. Hindsight being 20-20, clearly the sediment bed was over-characterized on a spatial basis (Fig. 5-5, p.5-12) relative to the spatial resolution adopted for the model (Fig. 5-6, p.5-13). However, the longitudinal structure is well characterized.

(*)f. The definition of TOC* needs to be clarified (p.5-11, lines 23-26); perhaps an equation?

(*)g. PCB partitioning to DOC two orders of magnitude weaker than for POC (p.5-14, lines 25/26); this needs literature support with respect to consistency (not found in Appendix B).

(*)h. What are the implications of not considering the partitioning between organic and inorganic particles within the four size classes, as it affects PCB associations? Does the Sediment f_{oc} fractionation (p.516, 5-17) procedure accommodate this? These protocols need clarification.

(*)i. The axis labels on Figure 5-11, left panel, and units ($\mu g/L_a$) have not been introduced in text first. The authors need to expand the description presentation of this section. **More than anywhere else in the report, this section fails to adequately communicate what was done and why it was appropriate**. Further, features of Figure 5-12 need to be better described - what is good or bad about these simulations.

j. It is impressive that parameters describing PCB partitioning were not adjusted as part of the calibration process.

k. PCB model performance for time series is not particularly impressive (note Log scale, Figures 5-16 through 5-18).

l. Longer-term validation modeling, covering more than 20 years, will follow this calibration effort and provide better basis for evaluation of performance for PCBs in sediment bed. This may mean calibration is not really complete, as some adjustments may be desirable to simulate the sediment record. Again, what version of the model will be used for such a long-term simulation, given the long run time issue.

m. The contrasting low and high Q longitudinal patterns and related performance are a positive feature, but there is substantial averaging embedded, as well as substantial variability in the observations (vertical bars; Figure 5-20 and 5-21). Times series performance was generally good, particularly for the May 19-21, 1999 event at New Lenox Road (Figure 5-23). Application of the model for Figure 5-25 is good.

(**)n. The modelers acknowledge shortcomings in certain features of predictions, in particular, "…. Measured concentrations shown in Figure 5-30 exhibit more variability than the model results …. Model cannot represent all the variability …." (P.5-58, lines 13-15). However, I take issue with the character of the explanation in the following sentence. It

appears to be an assertion. "Natural stochasticity"? What sources of variability in drivers were not represented that would cause these differences. I suggest the position be supported by model analyses. "Analytical variability" - is this the same thing as analytical errors. I suggest some representation of this be included; should be available from laboratory QA/QC records.

(**)o. The characteristics/features of performance presented in Table 5-4 are not treated until the end of the Summary of the section. This is inappropriate. Further, this performance needs to be treated in the wider context of other PCB-contaminated sites. This is particularly important for the concerned Members of the community. They deserve to know how this modeling Effort stacks up against others. My guess is that this performance will Rank fairly high. The comparison should be rigorous to the extent that is Reasonable (e.g., Tabular performance features).

p. Comprehensive sensitivity analyses were performed.

C. Bioaccumulation Modeling

1. Partitioning of biota appears to be a reasonable simplification and representation of the communities of concern.
2. Reasonable efforts were made to parameterize the model according to site-specific information - always an important feature in developing model credibility ("parameterization summary" - p.6.3).
3. Spatial segmentation, according to contrasting physical characteristics along the PSA, also seems appropriate.

(*)4. It remains somewhat unclear how supporting PCB fate and transport modeling is conducted for the 6-yr calibration (1995-2000) interval for FCM, particularly outside of the May 1999 - June 2000 interval (for Calibration of the PCB fate and transport model). Apparently the rest of the FCM calibration interval was "filled-in" with EFDC simulations for 1998 (p.6-7, lines 14-20). Exactly what are the implications of this approach? What are the benefits and effects of adding 1998 in for EFDC simulations to support FCM, instead of simply extending those for 1999-2000?

5. The specified contrasting conceptual models, according to reach (5A vs. 5C, 5D and 6), appear to be reasonable.
6. Why aren't the simplifying assumptions, conceptual models, etc., supported by references to the scientific literature, or at least similar gray literature?
7. The two-step calibration process, driven alternately by field data input vs. linked model inputs, is an attribute of the modeling effort.
8. Despite good efforts to constrain calibration parameters by site-specific measurements, there remains substantial degrees of freedom (e.g., knob-tuning) in the calibration process (section 6.3.3.2, p6-24 and 6-25).

(*)9. I am uncomfortable with the justification for averaging for performance evaluation for fish (lines 20-22, p.6.3.4). Find alternative language for "reduce the effect of these natural variations". Such data have been

observed to be inherently variable at contaminated sites. Use supporting literature citations.

(*)10. Apparently fish concentration predictions within a factor of **2** are accepted as good. This position should be supported by literature citations, rather than simply arguing that the project goal has been met.

11. Predictions of PCB concentration in invertebrates according to congeners was impressive (Figure 6-10, p.6-35): also generally good for fish (e.g., Figures 6-11 through 6-14).

12. Comprehensive sensitivity analyses were conducted (e.g., Tables 6-4 and 6-5).

13. Uncertainty analyses were documented that extend beyond the treatment given in EFDC modeling efforts. Is there a reason for the differences in treatment, and will efforts in this area converge in future modeling efforts for the site?

14. Probabilistic representations, as presented, are particularly appropriate to adopt for predicting responses of contaminant levels in biota to clean-up scenarios.

Attachment No. 2

Description of the Integrated Modeling Process

The Integrated Modeling Process

- critical roles of monitoring and process studies