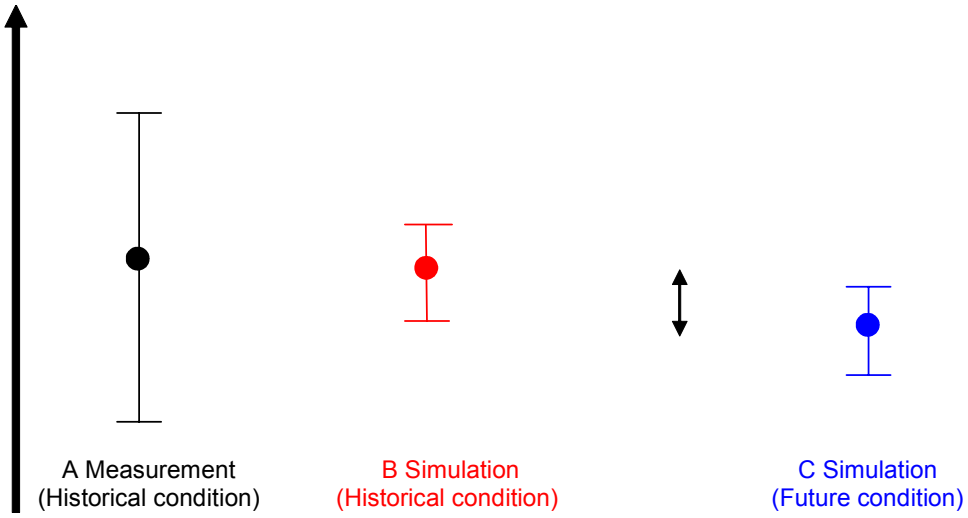
Final Peer Review Comments Regarding Housatonic Model Calibration

E. Adams

June 2005

1) Are the comparisons of the model predictions with empirical data sufficient to evaluate the capability of the model on the relevant spatial and temporal scales?

This is really the overarching question associated with model calibration: are the models good enough? But the question has to be asked in the framework of what the models will be asked to do, which is to evaluate remedial alternatives. (Problem identification is usually the first step in model evaluation; see, e.g., Ditmars et al., 1987.) The following sketch and discussion provide a framework for addressing this issue in general terms. Supplemental information is provided in the answers to the remaining five questions.



The black symbols on the left (A) represent, qualitatively, the range in space and time of measured state variables. The models have many state variables, but I will focus on three: flow rate; sediment concentration (bed load and suspended); and PCB concentrations (in the sediment, water column and fish). Some of the variables, such as sediment PCB concentrations, have tremendous variability as implied by the wide brackets. The mean is indicated by the circle. The red symbols in the center (B) represent the calibrated models' prediction of the same variables, while the blue symbols on the right (C) indicate the models' simulation of the same variables under a future remediation scenario.

We are being asked if, following calibration, the models are "good enough". In model calibration, this traditionally means asking how well the simulations (B) match the historical data (A) with respect to the mean, the variance, etc. However, it is more relevant to ask how confident we are that the models will correctly predict the

environmental effectiveness of various remediation options, which we are told include removing contaminated sediments by dredging and/or dry excavation, burying them through in situ capping above and/or below water, and natural recovery. Evaluation of future options introduces the following questions concerning model skill: how well does the model predict absolute mean output (an absolute measure indicated by the blue circle in C), how well does the model predict relative mean output (i.e., change from existing conditions, indicated by the black arrow representing the difference between the red circle of B and the blue circle of C), and how well does the model simulate the range in output (both in absolute terms and relative to the existing condition). In general a model will perform *better* in a relative assessment, than in a calibration, because model errors tend to cancel. Conversely a model may perform *worse* in an absolute assessment, than in a calibration, because different data and processes (reflecting remediation options) are involved. Of course we don't have measurements of future conditions with which to compare model results so we have to make inferences. This is done qualitatively below for the various types of model output.

Flow rates are primarily an output of HSPF and measured inputs, as filtered by river hydrodynamics (EFDC). HSPF has been around for a long time, its developers have had lots of experience with it, and the available data for the Housatonic seems to be on par with (or better than) what is typically available for other sites. There are a lot of semi-empirical parameters that can be adjusted to achieve a good fit and the fits displayed seem generally acceptable with respect to both mean and variance. (Some errors come when simulating storm events, but this is due to the difficult of getting storms right—both magnitude and timing—with only one hourly rain gauge in the watershed. But this should not be a problem in a statistical sense: if sediment-laden PCBs are being eroded due to a storm, we aren't concerned about the exact timing of the storm.) Because the proposed remediation measures should not significantly affect flows there is no reason to believe the model will not be able to adequately simulate flows under future conditions.

Sediment transport is performed by EFDC (with inputs from HSPF) and results from the processes of erosion, bed load transport (primarily of coarser size fractions), resuspension (primarily of finer size fractions) and deposition. There are no direct field measurements of bed load transport, but the model has been calibrated to produce reasonable agreement with field measurements of mean total suspended solids (TSS) suggesting that the net effect of the contributing processes is satisfactory, at least with respect to the mean. The agreement on the variance would best be seen using an equivalence plot of TSS (graph of measured TSS on the vertical axis and simulated TSS on the horizontal axis, with a 45 degree line indicating perfect agreement). Apparently such plots have not been generated, but I suspect that they would show that the model under predicts the variance in TSS for two reasons: First, TSS is a function of erosion, which depends on shear stress to a power n. Shear stress, in turn, depends on the local velocity squared, making erosion dependent on velocity raised to the power of 2n. Because EFDC uses approximately one grid cell per channel width, it can only output channel average velocities. Hence velocity extremes associated with lateral variations in channel depth and meandering are ignored. Furthermore, the calibrated values of n are approximately one, whereas W. Lick points out that the literature suggests the value

should be more like two. Hence a model calibrated to produce the right amount of resuspension on average will likely smooth out the extremes, in particular under predicting resuspension during high-energy events. It would likewise be expected to under predict the extremes in bed load transport. The under prediction of both bed load and suspended load transport would cause the model to underestimate the potential for natural remediation, while the underestimate of bed load transport could cause the model to underestimate the threat of cap erosion if instream capping were to be considered for remediation. I recommend that measured and predicted TSS be plotted on an "equivalence plot" to assess the magnitude of this problem and, if significant, the erosion model be recalibrated. Of course this means that deposition and possibly other processes would also need to be recalibrated.

PCB concentrations are predicted using both EFDC and the Food Chain Model (FCM)**.** EFDC has been calibrated to produce good agreement with measured average PCB concentrations in the water column. This means that the net effect of several exchange processes is in balance (though it doesn't say anything about the individual processes themselves). Over a short calibration period not much change can be expected in the sediment PCB concentrations so it is difficult to assess calibration here. The FCM seems to be doing a good job of reproducing at least mean concentrations in fish and EPA claims that this is all they care about. Remediation may result in lower sediment PCB concentrations (e.g. if some of the PCBs are removed) or a redistribution of PCB mass (e.g., if the PCBs are sequestered under a cap). To the extent that the sediment-water exchange processes, and the biological uptake, are linear with respect to concentration, good agreement between model and data under existing conditions should imply good ability to predict future conditions when the concentrations are lowered due to remediation. To the extent that the remediation removes a fixed fraction of both fine and coarse-grained sediments (and their associated PCBs), which seems reasonable for a dredging scenario, the assumption of linearity is reasonable. (See discussion in the following paragraph.) I agree with the concern expressed by QEA/GE that the depth of PCB bioavailability within the sediments (6 inches) is too great. Since PCB concentrations are currently fairly well mixed in the upper 6 inches, this does not affect the ability to simulate uptake presently, but it would affect the ability to predict uptake under a future scenario in which the PCBs were buried under cleaner sediments (either by application of a cap, or over time by natural processes).

During the Peer Review Meeting much was said about the tremendous spatial variability in sediment PCB concentrations over space scales of order one meter and the fact that the model can not reproduce this variability. The failure of the model to pick this up should not be considered model error, *per se*, but simply unresolved variability in model input and output (sediment bed concentration distributions). This variability is real and most likely reflects the stochastic method in which the PCBs were introduced in the first place. We cannot expect the model to predict this variability and the fact that the model averages concentration over relatively large grid cells is not a problem (with the mean) unless sediment-water exchange of PCBs varies non-linearly with concentration. Of course, we can not expect the model to tell us anything about the future variance of sediment bed concentrations, and to the extent this is important we should rely on the

observed variability. The PCBs have been in the sediments for several decades, and to a first approximation the variability expected in the next decade or two (presumably our focus) will not be very much different from the variability observed historically.

Back to the mean, sediment-water exchange *would* be expected to vary non-linearly with sediment concentration to the extent that the PCBs are associated preferentially with finer sediments that are more easily eroded. The model could be getting the flux of (primarily fine grained) sediments correct, but would be assigning an average PCB concentration to these sediments and hence underestimating the flux of PCBs. As a result I suspect the effect of averaging (sub-grid scale variability in sediment PCB concentrations) results in an underestimate of PCB flux during periods of high resuspension, for both existing and future conditions. To assess this effect, the correlation of PCB concentration and sediment type should be checked in available measurements and, to the extent possible in prediction. (I realize that model output is averaged over spatial scales that include a range of sediment type.) Also, although it would be a major change at this point, I wonder if it wouldn't be better to have the model formulated to predict sediment PCB concentrations simply as a function of sediment type (e.g., coarse, medium and fine) with only very coarse longitudinal discretization (say by reach 5A, 5B, etc.) This would result in much less model output (by two orders of magnitude), making the calculations more tractable, and the output would be more environmentally relevant: we don't care which 20 m cell within a reach a fish is in when it feeds in contaminated sediments, but merely the likelihood that the particular sediments will actually be contaminated—and to what extent.

2) Is there evidence of bias in the model, as indicated by the distribution of residuals as a function of the independent variables?

Comparison of predicted and measured mean values is reasonably good for most variables, as would be expected following calibration. But, as the discussion above implies, the models can be expected to under predict the fluxes of sediment and PCBs under extreme events, and these extremes will be most responsible for changes in PCB concentrations in the future. This is a type of bias.

3) Does the model, as calibrated, based upon your technical judgment, adequately account for the relevant processes affecting PCB fate, transport and bioaccumulation in the Housatonic River?

All three models are considered state of the art and hence come generally "fully equipped". The model team is experienced and has added important features such as the effects of vegetation on stream flow and transport. Hence I believe that most processes are at least *represented* in the models. I do agree with QEA/GE that bank erosion/river meandering is an important process that should be included and could help explain the large vertical spread of PCBs observed in cores.

Having said that, it is not clear that the models have been properly calibrated when it comes to individual processes. While each of the three models has their own unique calibration issues, EFDC is perhaps the most problematic because: 1) it is relatively new and has not been used in the current framework (e.g., with both in-channel flows and above bank flows), 2) it is being calibrated over a short period of time (14 months) relative to the time constants of some of the biochemical processes, and 3) compromises are being made because of computational expense. Including data from the earlier years (~1980-1999), as has been mentioned for the next phase, will help with the calibration. If this is done it would be nice if the last few years (2000-2004) could be set aside for a true validation (no more parameter tweaking).

One issue with calibration is that multiple variables contribute to a particular observation so the net effect may be correct, while the individual effect is not. As mentioned previously, one example is water column TSS concentrations that reflect both resuspension and deposition.

I am still a bit uncomfortable about the fact that most of the PCB mass is in the floodplain (and hence affected only by relatively rare flood flows), yet the vast majority of the computational time is taken with in-channel flows. Should the same model be used (in the same way) for both? It is not clear whether storms or the routine flows will be most responsible for PCB transport, but the fact that the model can only afford one grid cell over the channel width and a couple of grid cells in the vertical seems to defeat the purpose of a 3D model.

The mass transfer coefficient used to compute sediment-water flux is being calibrated to match observed changes in contaminant flux between two stations. Since the model has only one grid cell per river width, the calculated flux is based on a cross-sectional average flow rate, and hence an average velocity, bottom shear stress and erosion potential. It is quite possible that at a given time portions of a reach are eroding while others are depositing and it is not clear if a calculation based on average flow is correct. Furthermore, I would expect to have seen the mass transfer coefficient increase with river flow, but this apparently was not the case.

I am also concerned about the bioturbation coefficients. I realize that limited actual calibration has taken place so far, but I wonder if true calibration will be possible in the future. The tentative value of E-9 m2/s is quite large; are there biological observations to support such a value (or any value)? 14 months is too short a period to determine if bioturbation (plus diffusive flux across the interface) will have much effect; will adding an additional 10-20 years be that much more helpful, especially if the coefficient turns out to me much smaller?

Finally, the measured PCB sediment concentrations are not an ideal data set for calibration. As mentioned above, one problem is that there is not much change over (14 months of) time. There is some change over space, but this is overwhelmed by the much greater change over very short (sub-grid scale) lengths, which cannot be resolved. Finally, under historical conditions PCBs have been ubiquitous, appearing with the

upstream inflow, and eroding/diffusing off the sediment bed, river banks, and floodplains. These PCBs "all look the same" and hence it is hard to diagnose transport mechanisms based on a model's ability to "match" them. It was not discussed in detail at the Peer Review Meeting, but if there is time, serious consideration should be given to a tracer experiment of some sort. The best type of experiment, in theory, would be one in which sediment of different types (or placed in different locations) were uniquely labeled (e.g., with fluorescent colors) so that they could be tracked. This would take time and resources to think through, but it may be the only way to calibrate certain model parameters. An alternative approach would be to conduct a trial remediation on a patch of river and see how rapidly contaminated sediment from upstream fills in the clean spots. (In principle one could simulate the effect of clean upstream sediments that will result from ongoing remediation to see how fast they fill in the contaminated portions downstream, but it is better to simulate a clean downstream spot since a bit of contamination in an otherwise clean patch shows up more than a bit of clean sediment in an otherwise dirty patch, especially given the variability.)

4) Based on your technical judgment, have adequate methodologies been employed to evaluate the sensitivity of the model to descriptions of the relevant processes, and to evaluate uncertainties of model predictions?

The analysis of sensitivity and uncertainty has been performed in the context of historical conditions; it should be repeated for future remedial conditions. That is, the model should be set up to assess the effectiveness of different remediation measures, and the sensitivity of this effectiveness to various model parameters assessed. This may lead to quite different conclusions. In some cases is might also be a lot easier. For example, if dredging is being considered, it is easy to simulate the effect in the model: the PCBs will simply be removed and the effectiveness of this remedial option (at least in a relative sense) will be totally insensitive to most model parameters. (There would be some short term disturbances during the actual dredging operation, but these would need to be simulated by different models, for which the current calibration study is irrelevant. Such disturbances are often considered part of an environmental impact assessment, but have not been mentioned; are they something to be considered?) Another example is capping which will introduce strong vertical gradients in PCB concentration within the sediments. Under this option the model may become much more sensitive to the way in which mixing between sediment layers is represented. Indeed, it may not be possible to calibrate some parameters based on historical field data (e.g., because the gradients are insufficient) so recourse must be made to literature values.

A major difficulty with sensitivity and uncertainty analysis is that the model package is very expensive to run (owing principally to the long run time of EFDC reflecting, in turn, the small time step), and this problem will only get worse as the length of the simulation increases (from 14 months to order 20 years). Here I think the consultants could have tried to "think outside the box" a bit more with respect to the way the hydrodynamics, sediment transport and water quality were coupled. For example, the hydrodynamics could be run off-line and saved. And perhaps synthetic hydrologic sequences could be

used. Or different grids could be used for the in-bank versus over-bank flows. QEA/GE made several suggestions along these lines at the Peer Review Meeting. If there is time, these should be seriously considered as a way to decrease computation time and allow the modeling team to afford to be able to run additional calculations.

5) Is the uncertainty indicated by model-data differences sufficiently inconsequential to permit use of the model to predict differences in remedial options?

See response to Question 4.

6) Are the processes in the model calibrated to the extent necessary for predicting future conditions including future concentrations of PCBs in the environment under natural processes and under potential remedial options for sediments and floodplain soils in the Housatonic River in the reach below the confluence? If not, what additional work needs to be done to calibrate the model?

See response to Questions 1, 2 and 4 which also include suggestions for improving model calibration.

Reference

Ditmars, J.D, E.E. Adams, K.W. Bedford and D.E. Ford (1987). "Performance evaluation of surface water transport and dispersion models" *Journal of Hydraulic Engineering, ASCE*, 113(8): 961-980.