Housatonic River Rest of River Ecological Risk Assessment Review Panel Comment Submission Form – Final Comments

Name of Panel Member: Valery Forbes
Date: 29 January 2004

Executive Summary - Overall Recommendations for Improving the Risk Assessment

1. The assessment endpoints should be redefined so that they are more consistent with general EPA practice and so that they more accurately reflect the protection goals that were actually used in this ecological risk assessment (i.e., long-term persistence of local receptor populations).

2. More transparency and consistency is needed in describing the WOE approach. Describing the process, or parts of it, using the phrase 'best professional judgement' should be avoided. More care should be taken in combining lines of evidence that are not independent. The WOE summary tables should be modified so that they are more self-explanatory and less ambiguous.

3. More detailed and consistent descriptions of the statistical methods used should be provided in those parts of the ERA where data are presented (the reader should not be referred to the original article to find out what kind of statistical test was used). Both statistical significance and effect size should be reported and considered in the risk characterization.

4. Interpretation of HQ results needs to be refined. Both the magnitude of the maximum HQ as well as a measure of the probability (or proportion of samples) exceeding an HQ of 1 (or 10, or 100 as appropriate) should be included; it should be clear whether the spread in the HQs derives from variability in exposure (the numerator), variability in effects (the denominator), or both. Given that HQs provide a rather coarse measure of risk, differences in HQs of less than an order of magnitude should not be considered as indicating differences in risk.

5. The ERA should avoid use of value-laden terms to describe risk (e.g., catastrophic, unacceptable), and instead aim to quantify the likelihood and degree of impact in objective terms as best as possible.

6. The panel identified a number of studies/analyses that could have been done in the context of the risk assessment. I do not recommend that completion of the ERA be delayed in order to include more studies in it. However, given that an important output of the ERA is the identification and quantification of important sources of uncertainty, I would strongly recommend that actions taken on the basis of the ERA include both consideration of remediation alternatives as well as additional, highly focussed, studies/analyses designed to address the most important uncertainties identified in the ERA.

7. Serious consideration should be given to restructuring the ERA to limit the redundancy between the Assessment Endpoint Chapters in the main document and the relevant Appendices in which all of the details are found. In my view the Endpoint

Chapters provide too much information for the casual reader and not enough for the interested expert. These could be deleted from the main document since all of the information they contain is provided in the Appendices. A series of maps that overlay sampling sites for exposure estimates and sampling sites for the various effects estimates would be a very helpful addition to the document.

8. According to EPA guidance, ERAs should use site specific studies wherever possible. Unfortunately many of the field studies performed in the context of the present ERA suffered from weaknesses related to one or more of the following: no reference sites; small sample sizes; short study durations (e.g., one reproductive season); they addressed a question that did not lend itself easily to incorporation in the WOE (e.g., is species X reproducing in the PSA, yes or no?). This is extremely unfortunate since the potential strength of site specific field studies is that they deal directly with mixtures of chemicals (and other stressors) present at the study site and should therefore have less uncertainty (and weigh more heavily) than laboratory studies or models. I would recommend that EPA and GE work together toward developing some guidance on the appropriate design of field studies for use in these kinds of ERAs in the interest of improving future projects of this nature.

9. It would be extremely valuable if the EPA and GE could jointly compile a document that highlights the lessons learned from the Housatonic risk assessment project in a format that could provide guidance for the successful conduct of future risk assessments of this kind.

Detailed Answers to the Charge Questions

My answers to the Charge questions are based primarily on the main ERA but include, where relevant, EPA's responses to Panelists' written questions and oral responses provided at the public meeting held 13-16 January 2004. Thus I am assuming that if the requested information was not present in the main ERA but was addressed satisfactorily in the EPA's written or oral responses that appropriate amendments will be made following the Peer Review meeting.

Charge Question 1. Was the ecosystem of the Housatonic River watershed properly characterized, and was this information appropriately applied in the Problem Formulation and subsequently in the ERA?

Comments: The ecological characterization seems to have been extremely thorough, and a relatively detailed knowledge of the ecology and habitat usage, particularly of the birds and mammals, seems to have been incorporated into the ERA. However I feel it is unsatisfactory that the assessment endpoints were chosen, to some extent, on the basis of whether or not data were available for the species under consideration (EPA response to Panel Question BS1). I would argue that the availability of data is not an appropriate criterion for selection of assessment endpoints (though it can be a constraint for selecting measurement endpoints). If there is an endpoint for which protection is deemed an appropriate goal on the basis of the site characterization, then the necessary data should be collected as part of the ERA.

Proposed Changes: A detailed road map or data inventory could increase clarity and reader-friendliness. A figure (or series of figures) showing spatial variation of tissue sample sites and concentrations could be a useful addition.

The ERA should include an explanation of why some of the risk characterization studies were not included in the ERA (e.g., dragon flies, mussel, blue gills). Better overviews (tables or figues) of what data have been used would improve the document.

Charge Question 2. Was the screening of contaminants of potential concern (COPCs), selection of assessment and measurement endpoints, and the study designs for these endpoints appropriate under the evaluation criteria?

Comments: The screening of COPCs was generally appropriate. The use of the pre-ERA to identify COPCs other than PCBs and to determine the downstream boundary beyond which PCBs from the GE facility pose a negligible risk to aquatic biota and wildlife was an effective approach. Nomenclature concerns (Panel Question BS2) could be addressed by referring to the pre-ERA as the Initial Risk Assessment and the ERA as a Refined Risk Assessment. Also, the 3-step tiered approach for establishing an initial COPC list seems to be appropriately conservative with the possible exception of Tier 3 in which evaluation was performed 'subjectively'.

From p. 2-58 the assessment endpoints are defined as representing 'specific ecological values deemed important to protect', whereas measurement endpoints are defined as 'the tools used to determine the outcome for the assessment endpoints'. Although it is possible that some measurement endpoints may also be assessment endpoints, in my view the assessment endpoints defined in this ERA (with the exception of community structure) would be more appropriate as measurement endpoints whereas the assessment endpoints would be more appropriately defined as the long-term persistence of populations of benthos, fish, amphibians, birds and mammals in the PSA. To some extent the defined assessment endpoints are redundant. For example, changes in benthic community structure occur because of changes in survival, growth, and/or reproduction of resident species. This is reflected in the WOE for the benthos which states 'the individual measurement endpoints were often applicable to many or all of the assessment endpoints' (D 94) and thus a single WOE was performed that included all benthic assessment endpoints. However, if the assessment endpoints are as stated then benthic toxicity results using different responses (e.g., mortality versus reproduction) should, in principle, have been analysed separately (since they represent separate assessment endpoints) instead of being put into the same analysis. This is probably an issue for other receptors as well.

As is stated by EPA (response to Panel QuestionJO7), 'Any contaminant-induced response that leads to direct mortality of adult fish, and/or indirect effects on population structure (e.g., loss of recruitment of juveniles to older age classes), and/or health (e.g., reduction in fish growth rates, reduced adult reproduction rates*) that lead to an impact on the locally-exposed population* [emphasis added] would be considered an ecologically significant response.' This suggests that populations were, in effect, the objects of protection in the present ERA.

I can further point out that populations are specifically named as targets of protection by EPA (1998). When the focus is on the population as a whole, it is acknowledged that a stressor may affect the survival, growth and/or reproduction of some members of the population but that the "acceptability" of the stress is judged in terms of how it effects the population as a whole.

A practical problem with the assessment endpoints as defined is that having several assessment endpoints for each receptor forces the assessor to make judgements as to whether, for example, reproduction, survival, development, maturation, and community condition of amphibians are of equal importance, if the most sensitive of these should drive the risk characterization, or if some should be given more importance than others. An example is for bald eagles where the risk of TEQ was determined to be high for eggs, but low for adults, and the WOE concluded an intermediate risk. Depending on the life-history characteristics of the species, the survival of eggs versus adults may differ in demographic importance. In addition, it is incorrect to assume that high risks for individual performance indicators necessarily and consistently translate into high risks for the population. Clearly EPA recognizes this (see e.g., response to Panel Question MAO2), but have not made the link quantitative. One ecologically based way to weigh risks to different life stages is to consider their importance in terms of population dynamics (e.g., by an elasticity analysis).

For threatened and endangered species the individual is often defined as the protection goal. Partly this is because loss of any or few individuals may have a measurable influence on the population's persistence. However for most other taxa considered, it is persistence of populations, and not individuals, that is the protection goal. Indeed, on page 2-66 it is stated that 'Although many of the endpoints presented are linked to organism-level effects (e.g., survival and reproduction), these endpoints are expected to be strong indicators of potential local population-level effects'. While this is broadly true, the form of the relationships between organism-level effects and population-level effects will vary widely among endpoints and species. Organism-level effects can act as measurement endpoints for estimating population-level effects, but the links should be made quantitative (e.g., through demographic or life-cycle models).

Proposed Changes: I would propose that serious consideration be given to redefining the assessment endpoints: reproduction, growth, and survival as measurement endpoints for the target species considered, and that the assessment endpoints be redefined as 'long-term persistence of populations of receptors'. Likewise it should be clear that for example 'amphibians' are a receptor, whereas Leopard and Wood Frogs are surrogate species chosen to represent amphibians. Also for the other receptors.

Charge Question 3. For each of the 8 assessment endpoints evaluated in the ERA (listed in Attachment B, and for which a specific Section and Appendix was prepared), address the following questions (discuss and label responses as 3.(assessment endpoint number).(question letter) for consistency):

3.1 Benthic Invertebrates

(3.1.a) Were the EPA studies and analyses performed (e.g., field studies, site-specific toxicity studies, comparison of exposure and effects) appropriate under the evaluation criteria, and based on accepted scientific practices?

The sediment quality triad approach is a potentially powerful one for assessing risks to benthic communities. Environment Canada has developed a very useful guide to interpreting results of triad assessments, particularly when the different lines of evidence give conflicting conclusions (Reynoldson et al. 2002, HERA 8:1569-1584). There are also other relevant papers in this special HERA issue (2002, volume 8, no. 7) on WOE in sediment risk assessment.

(3.1.b) Were the GE studies and analyses performed outside of the framework of the ERA and EPA review (e.g., field studies) appropriate under the evaluation criteria, based on accepted scientific practices, and incorporated appropriately in the ERA?

No GE studies performed. GE's reanalysis of benthic community structure is a relevant contribution and should be incorporated.

(3.1.c) Were the estimates of exposure appropriate under the evaluation criteria, and was the refinement of analyses for the contaminants of concern (COCs) for each assessment appropriate?

Given the extremely high spatial and temporal variability in sediment PCB concentrations (and to some extent other COCs), it is unfortunate that a number of the chemical measurements could not be easily matched with toxicity and/or community structure information. The difference in sediment concentration trends (stations 4 – 8) between the benthic community samples (sediment PCB concentration declines) and the toxicity station samples (sediment PCB concentration increases) is unfortunate and does not increase the clarity of interpretation.

The laboratory toxicity tests should use the most synoptic sediment concentrations for estimating exposure whereas for field community structure it is possible to include paired sediment concentrations from same sites/samples.

(3.1.d) Were the effects metrics that were identified and used appropriate under the evaluation criteria?

I question the use of *Daphnia* and *Ceriodaphnia* as appropriate benthic invertebrate test species. It would have been better to use another infaunal or epifaunal temperate invertebrate.

With regard to differences in the relationship between taxonomic diversity and sediment PCB in fine- versus coarse-grained habitats, it could be that the substrate difference is explained by differences in taxonomic composition between fine and coarse sites or that there are differences in PCB bioavailability (e.g., less bioavailable in fine-grained sediments) that could explain these differences. Sampling of benthos in the field differed somewhat for upstream coarse grained (wading in shallow water) versus downstream fine-grained (from boat with fauna collected along shore therefore larger spatial separation in latter 10-20 m). Whereas this may have been unavoidable, the differences should be mentioned in the discussion of fine- vs. coarse grained site differences.

It seems that the MATCs are ultimately based on only two species with multiple (non-independent) response endpoints, and this should be rectified. With regard to deriving MATCs, it is recommended that acute and chronic test endpoints be separated, that only one endpoint be used per species (could be lowest or could be geometric mean), that only the most synoptic data are used as measures of exposure, that only those tests that displayed a clear concentration-response relationship be used, that only sediment-relevant test species be used, that all of the available test species be used (i.e., not just the lowest 6 values), and that if the derived MATC is equal to or lower than the concentration at reference sites the value should be truncated at the reference concentration.

(3.1.e) Were the statistical techniques used clearly described, appropriate, and properly applied for the objectives of the analysis?

The statistical methods seem generally appropriate. However, the ERA could benefit from a better description of the statistical methods used. Enough detail should be presented so that the analyses could be repeated.

Shannon-Wiener may not be best measure of diversity for the sediments in which a few species dominate (Tom La Point suggested Simpson's index).

I believe that the concerns raised by GE in response to the reanalysis of the benthic data are important. If a small fraction of the total variability in benthic species abundance can be explained by PCB concentration, despite statistical significance of the regression, this suggests that the role of PCBs in determining benthic community structure may be less important than concluded by EPA.

I recommend that both effect size and significance are important and should be presented for all experimental results where appropriate. This is true throughout the ERA.

(3.1.f) Was the characterization of risk supported by the available information, and was the characterization appropriate under the evaluation criteria?

Regarding the multiple regression analysis provided in response to Panel's questions it would seem that the role of PCBs as a major factor influencing the abundance of benthic invertebrates is questionable. Both proportion of variance explained as well as statistical significance need to be taken into account in interpreting these analyses.

The risk terminology used to describe HQs (i.e., definitions of low, moderate and high risk) needs checking for consistency with other COCs as well as with other assessment endpoints throughout the ERA. HQs should be used as rough estimates of relative risk within assessment endpoints. Broad brush order of magnitude differences could be useful indicators of relative risk. Other COCs have HQs greater than one but the contribution of these was downplayed. See figure 4.2. There is a need for greater consistency in the interpretation of HQs exceeding one. Also the magnitude and frequency of exceeding the relevant threshold should be considered. It is essential to point out that for PCBs variability in the HQs reflects variability in the exposure estimates, with a single value representing the effects. For other COCs HQ variability reflects variability in the effects thresholds with a single point estimate for exposure.

(3.1.g) Were the significant uncertainties in the analysis of the assessment endpoints identified and adequately addressed? If not, summarize what improvements could be made.

The uncertainties in linking sediment chemistry to toxicity and community structure were largely addressed by analyzing different subsets of the available data (e.g., most synoptic, median). This was a useful approach. However I found very confusing the presentation of the sediment chemistry data for the toxicity and community structure samples plotted by station as it required careful reading (and explanation by EPA) to clarify that these chemical concentrations were not necessarily representative of the stations.

It should be emphasized here that a substantial fraction of the 'uncertainty' is actually true variability in exposure of benthic receptor species. Such variability cannot be reduced by further measurements and should be interpreted differently in assessing risk than uncertainty due to lack of knowledge.

(3.1.h) Was the weight of evidence analysis appropriate under the evaluation criteria? If not, how could it be improved?

As stated on p. 2-66, 'no matter what form the WOE takes, it should provide documentation of the thought process used when assessing potential ecological risk'. The weights are determined on the basis of 10 attributes that reflect the strength of association between assessment and measurement endpoints, data and study quality, and study design and execution. It is unclear how the total value for each measurement endpoint is achieved from the scores of the 10 individual attributes (e.g., Fig 2.9-1). According to the EPA's response to Panel Question VF16, the 10 attributes were considered of equal importance and the total endpoint values were determined using best professional judgement based upon the values assigned for each of the attributes. The ERA would be much more transparent if the best professional judgements were articulated more clearly.

I cannot find a description of how the overall assessment within a measurement endpoint is determined. For example how are the symbols in the right-hand column of Table D 3.3 determined from the combinations of symbols for the different toxicity test results?

The inclusion of different numbers of effects endpoints for different species can potentially bias the WOE. For example if a species that is either very sensitive or very tolerant has more measurement endpoints than other species going into the analysis, this can lead to a biased assessment.

Likewise when the data are scored for evidence of harm and magnitude, it seems illogical to have scores for magnitude in the event that evidence of harm is either 'no' or 'undetermined'. In EPA's response to this question (Question VF14), it is explained how such a combination of scores might be possible. This explanation should be included in section 2. Nevertheless, there must be some combinations that cannot logically occur. To follow the EPA's example, if a field study could not rule out high risk, it would be illogical to conclude 'undetermined/high', because the risk could just as well be intermediate or low.

(3.1.i) Were the risk estimates objectively and appropriately derived for reaches of the river where site-specific studies were not conducted?

The general approach of selecting target groups based on risks observed in the PSA and downstream occurrence of the target species in combination with mapping of threshold concentrations seems logical and cost-effective. However, there seems to be some public concerns that the CT portion of the river may not have been adequately assessed. It would seem that with relatively little effort and expense, additional sediment samples could be analyzed from CT portions of the river (as recommended by Peter DeFur) which could go a long way toward alleviating these concerns and strengthening the conclusions of the risk assessment. These could be taken as one of the 'management actions' taken on the basis of the ERA.

(3.1.j) In the Panel members' opinion, based upon the information provided in the ERA, does the evaluation support the conclusions regarding risk to local populations of ecological receptors?

The ERA concluded that risk is high for benthic invertebrates and that confidence in this conclusion is also high. In my view the benthic invertebrate data are more equivocal than indicated in the ERA. This is largely due to the substantial spatial and temporal variability in sediment PCB concentrations and the rather surprising (to me) difference in the relationship of taxonomic diversity versus PCB concentration between coarse and fine-grained sediments. The potential contribution of other COCs needs further attention (check especially for consistency in interpretation of HQs). One approach could be to do a multivariate analysis including other COCs. A re-ananalysis of the community structure data is warranted. HQs could be re-assessed as frequency exceeding the threshold. Dose-response relationships of toxicity data using most synoptic chemistry data need checking. In addition, consideration should be given to including dragonfly data, crayfish data and any other relevant data from the risk characterization that have not been included.

3.2 Amphibians

(3.2.a) Were the EPA studies and analyses performed (e.g., field studies, site-specific toxicity studies, comparison of exposure and effects) appropriate under the evaluation criteria, and based on accepted scientific practices?

Generally yes. In principle I believe it could be efficient to use some of the field studies performed for site characterization in the risk assessment (e.g., vernal pool surveys for breeding amphibians Appendix A.1). Unfortunately these were concluded to be an insensitive tool for detecting effects of PCBs.

As stated above, I believe that the definition of assessment endpoints for amphibians is inappropriate.

The design of both the leopard frog and wood frog site-specific toxicity tests (FEL 2002) was rather involved and therefore somewhat difficult to follow. In both studies an excellent gradient of sediment PCB concentrations in the test pools was achieved. However, it was determined that exposure of egg masses and young was largely via maternal transfer and not pool sediment which, to some extent, complicates interpretation of the early life stage results.

In the site-specific toxicity study of leopard frog reproductive success, it was a weakness that no frogs were captured from the reference area and that the study had to rely on purchased frogs for the control group. Thus, the reference group is not a true control and should be dropped from the statistical comparisons. In this same study there were found low stage VI oocytes at all stations which was suggested could be due to frogs moving among sites (questioning actual exposure-response relationships). There was also a very small sample size available with only one to a few egg masses collected per pond.

(3.2.b) Were the GE studies and analyses performed outside of the framework of the ERA and EPA review (e.g., field studies) appropriate under the evaluation criteria, based on accepted scientific practices, and incorporated appropriately in the ERA?

Although I am not an expert in amphibian field studies it seems that the field studies performed here (i.e., leopard frog egg mass surveys) were not particularly powerful tests of potential PCB effects on frog populations due to problems linking actual exposure to observed effects and to small sample size.

The wood frog study by Resetarits (2002) seems to have been well designed (i.e., randomized complete block design, large numbers of larvae per treatment), but did not adequately simulate exposure of frogs to PCBs in the field (i.e., which would include both maternal transfer and sediment exposure).

(3.2.c) Were the estimates of exposure appropriate under the evaluation criteria, and was the refinement of analyses for the contaminants of concern (COCs) for each assessment appropriate?

Some uncertainties in exposure in some of the field studies as indicated above. No issues with COCs.

(3.2.d) Were the effects metrics that were identified and used appropriate under the evaluation criteria?

The relationships between metamorph malformations, sex ratio and population-level effects were not quantified which makes interpretation of the seriousness of effects on the measured endpoints difficult.

Also see points on derivation of MATCs for invertebrates.

(3.2.e) Were the statistical techniques used clearly described, appropriate, and properly applied for the objectives of the analysis?

Generally yes. The exception here is with EPA's leopard frog study in which the control (composed of purchased frogs) was not a true statistical control.

(3.2.f) Was the characterization of risk supported by the available information, and was the characterization appropriate under the evaluation criteria?

In my view applying a population modelling approach to integrate effects of PCBs (and other potential stressors, habitat features, etc.) on the individual-level endpoints measured can add considerable strength to the risk assessment. Such models can be particularly useful, for example, for comparing impacts on different life stages (e.g., how much of an impact on egg production would be equivalent to a given effect on adult mortality in terms of population-level impact?). Such an approach could have been applied to the other receptor species, especially where the different assessment endpoints showed non-congruent response patterns.

As far as I can determine, given the way that the input parameters were chosen for the model used here, the addition of PCBs would have to increase the probability of extinction (unless the increased larval survival with PCB exposure could offset all of the modelled negative impacts). So although I was not surprised to see that the PCB cases increased the probability of decline I find myself asking, 'but how much of an increase in probability of decline is too much?'. I also found it intriguing (and non-intuitive) that if the modelled frog population was

already declining, the additional impact of PCBs seemed to be less than if the population started from a stable state.

I recommend that the model be further explored, including consideration of various scenarios as well as a sensitivity analysis of model parameters.

(3.2.g) Were the significant uncertainties in the analysis of the assessment endpoints identified and adequately addressed? If not, summarize what improvements could be made.

The best way to address the uncertainties indicated in the field studies (due to small sample size and lack of information on actual exposure) would be to perform additional studies.

(3.2.h) Was the weight of evidence analysis appropriate under the evaluation criteria? If not, how could it be improved?

Sections 4.7.1.1 – 4.7.1.3 were excellent – a clear and transparent description of the thought process going into the weighting criteria.

Apparently GE's wood frog study measured 11 endpoints but only found effects on 2 (malformations and sex ratio). However the ERA only focused on the 2 that showed effects, despite that other of the endpoints are relevant for assessing survival and reproduction. These other endpoints should be incorporated into the WOE.

(3.2.i) Were the risk estimates objectively and appropriately derived for reaches of the river where site-specific studies were not conducted?

Yes, the landscape analysis in combination with sediment PCB concentrations seems to be a good way to do this. It is unfortunate however that there were no sediment samples available from the downstream vernal pool habitats. Taking such samples would be one way to reduce uncertainty.

(3.2.j) In the Panel members' opinion, based upon the information provided in the ERA, does the evaluation support the conclusions regarding risk to local populations of ecological receptors?

The ERA concluded that risk to amphibians is high and that confidence in this conclusion is high. Although I agree that the probability of some effects occurring in amphibians is high, it is not as clear to me that the magnitude of these effects is high.

3.3 Fish

(3.3.a) Were the EPA studies and analyses performed (e.g., field studies, site-specific toxicity studies, comparison of exposure and effects) appropriate under the evaluation criteria, and based on accepted scientific practices?

Neither the EPA nor the GE field studies were optimally designed to test concentration-response relationships. However both studies seemed appropriate for assessing the condition of fish populations in the PSA and therefore contribute important information.

(3.3.b) Were the GE studies and analyses performed outside of the framework of the ERA and EPA review (e.g., field studies) appropriate under the evaluation criteria, based on accepted scientific practices, and incorporated appropriately in the ERA?

See response to 3.3.a.

(3.3.c) Were the estimates of exposure appropriate under the evaluation criteria, and was the refinement of analyses for the contaminants of concern (COCs) for each assessment appropriate?

Mapping of exposure of fish populations in space would be a very useful addition; i.e., where were fish tissue data collected?. However, it is recognized that for some COCs fish tissue would not be a good measure of exposure.

(3.3.d) Were the effects metrics that were identified and used appropriate under the evaluation criteria?

The measurement endpoints used in the Phase I and II toxicity studies were appropriate, however linking them to impacts on fish populations is more problematic.

Some of swim bladder abnormalities seem to disappear with age. This issue needs further consideration.

Phase I spawn success data (number of spawns evaluated for abnormalities) have small sample sizes; and no clear dose-response. I recommend including only effects that show a dose-response.

In general, care needs to be taken when basing effects estimates on the surviving portion of the population especially if survival was very low and/or variable among treatments.

(3.3.e) Were the statistical techniques used clearly described, appropriate, and properly applied for the objectives of the analysis?

More details on the statistical methods are needed.

(3.3.f) Was the characterization of risk supported by the available information, and was the characterization appropriate under the evaluation criteria?

It is my understanding that some of the deformities observed in the Phase I toxicity study (USGS) are also consistent with Hg and/or PAH toxicity. I did not see this reflected in Appendix F.

The conclusion of the assessment was ‘low risk’ despite evidence of impairment with respect to the assessment endpoints. Justification (EPA response to Panel Question JO34) is that ‘the magnitude of that harm appears to be sufficiently low as to not result in observed population-level effects’. Again this would indicate that it is persistence of fish populations that is the actual assessment endpoint being employed.

The bias of field populations toward older individuals should be further considered for other possible explanations than lack of fishing.

(3.3.g) Were the significant uncertainties in the analysis of the assessment endpoints identified and adequately addressed? If not, summarize what improvements could be made.

On the basis of the site-specific toxicity tests evidence/magnitude of harm to fish (and their associated uncertainties) were presented as hazard quotients, with the variation reflecting only variation in the numerator (i.e., in fish tissue concentrations). Hazard quotients were also used for other receptors (e.g., benthic invertebrates) but I am not entirely sure whether the size of the HQ can/should be interpreted the same for all receptors given that the effects thresholds are based on different kinds of tests. Cumulative plots of tissue burdens that show probability of exceeding thresholds (sent in response to written question from JO) are a useful way to present the data. Other COCs with HQs exceeding one should be discussed.

(3.3.h) Was the weight of evidence analysis appropriate under the evaluation criteria? If not, how could it be improved?

Generally, yes. However, I feel that the field work (p. 5-70) was underweighted. Also the largemouth bass (GE) field study seems to indicate no evidence of harm. EPA claims this study shows undetermined harm. EPA's conclusion either requires further justification or needs to be changed.

(3.3.i) Were the risk estimates objectively and appropriately derived for reaches of the river where site-specific studies were not conducted?

The factor of 4 used to extrapolate risk to trout seems somewhat arbitrary. I noted in response to Panel Question JO32 that the EPA responded to the effect that the factor of 4 applied to the fish toxicity was not a safety factor but rather an extrapolation factor to account for potential sensitivity differences. Extrapolation factors and safety factors are actually synonymous (as are application factors and uncertainty factors). In any case justification for the use of the factor 4 in this instance to account for sensitivity differences should be spelled out more clearly in the ERA.

(3.3.j) In the Panel members' opinion, based upon the information provided in the ERA, does the evaluation support the conclusions regarding risk to local populations of ecological receptors?

The ERA concluded that risk is low to moderate for fish and that confidence in this conclusion is moderate. I agree that risks of PCBs for the persistence of fish populations are low, but I believe that confidence in this conclusion is high. The extent to which PCBs cause morphological deformities in individual fish is a separate issue, but clearly one of some interest to the public.

3.4 Insectivorous Birds

(3.4.a) Were the EPA studies and analyses performed (e.g., field studies, site-specific toxicity studies, comparison of exposure and effects) appropriate under the evaluation criteria, and based on accepted scientific practices?

Yes, the tree swallow study seems to have been appropriately conducted. However the fact that the tree swallow seems to be relatively insensitive to PCBs causes some concern and perhaps needs to be better argued for in the ERA.

(3.4.b) Were the GE studies and analyses performed outside of the framework of the ERA and EPA review (e.g., field studies) appropriate under the evaluation criteria, based on accepted scientific practices, and incorporated appropriately in the ERA?

Yes, the American robin study was well conducted and is one of the few studies included in the ERA which provided a formal statistical power analysis (provided in the ET& C 2003 publication).

(3.4.c) Were the estimates of exposure appropriate under the evaluation criteria, and was the refinement of analyses for the contaminants of concern (COCs) for each assessment appropriate?

Yes, in general. The probabilistic exposure models used for insectivorous birds and the other wildlife species are state of the art. Although one could disagree with some of the details of the input parameters, they are all clearly expressed and presented in a nicely organized way.

I question the use of energy content of grasshoppers and crickets to represent emergent aquatic insects for the tree swallow and robin assessments. It would not have been that difficult to make measurements of more appropriate prey species as part of the ERA.

(3.4.d) Were the effects metrics that were identified and used appropriate under the evaluation criteria?

Yes, the combination of literature-based effects thresholds and field study results was appropriate.

(3.4.e) Were the statistical techniques used clearly described, appropriate, and properly applied for the objectives of the analysis?

Yes, well done.

(3.4.f) Was the characterization of risk supported by the available information, and was the characterization appropriate under the evaluation criteria?

For the wildlife in general the risk characterizations were the most quantitative. For example, defining risk in terms of ‘the probability of a certain percent effect occurring’ clearly has advantages over hazard quotients.

(3.4.g) Were the significant uncertainties in the analysis of the assessment endpoints identified and adequately addressed? If not, summarize what improvements could be made.

Yes. Possibly with the exception of using a point estimate for concentration of PCBs in prey. Tree swallow HQs were not based on site specific data that could have been used, but rather on the basis of modelled data. I would recommend using appropriate site-specific data whenever possible to reduce uncertainty.

(3.4.h) Was the weight of evidence analysis appropriate under the evaluation criteria? If not, how could it be improved?

The total value of each measurement endpoint is described as a weighted average of the 10 attributes going into the WOE (p. G 77). Correction needed. This was not a weighted average. I believe that earlier chapters in the ERA did not describe the process by which the total values were determined. This should be remedied.

My assessment of the American Robin field study (at least the version published in ET&C 2003) was more positive than EPA's and I would have given it somewhat more weight in the assessment than was done.

Once again, I find the WOE scoring sheets can be ambiguous. A combination of 'no evidence' for a 'low magnitude' of harm (i.e., both field studies) could either be interpreted that the study would have been able to detect a low magnitude of harm but didn't or that the study did not find evidence for harm but there nevertheless could have been low magnitude harm that was not picked up. Reconsideration should be given to the table inputs so that they can stand alone and be interpreted without ambiguity.

(3.4.i) Were the risk estimates objectively and appropriately derived for reaches of the river where site-specific studies were not conducted?

Extrapolation to other species seems to have been performed as a qualitative comparison that only incorporated differences in exposure. It would be helpful if some argumentation could be provided to indicate whether these other species are likely to be more/less/ or equally sensitive to PCBs. Because low risk to insectivorous birds was indicated in the PSA, no risk assessment was performed for birds outside of the PSA.

(3.4.j) In the Panel members' opinion, based upon the information provided in the ERA, does the evaluation support the conclusions regarding risk to local populations of ecological receptors?

The ERA concluded that risk to insectivorous birds is low, but that confidence in this conclusion is not high (does that mean low or moderate?). I concur that the risks to this receptor are low, and I would put moderate confidence on this conclusion, largely on the basis of the field studies.

3.5 Piscivorous Birds

3.5.(a) Were the EPA studies and analyses performed (e.g., field studies, site-specific toxicity studies, comparison of exposure and effects) appropriate under the evaluation criteria, and based on accepted scientific practices?

The studies basically consisted of modelling, and yes these were appropriate.

(3.5.b) Were the GE studies and analyses performed outside of the framework of the ERA and EPA review (e.g., field studies) appropriate under the evaluation criteria, based on accepted scientific practices, and incorporated appropriately in the ERA?

The belted kingfisher field study was unfortunately conducted for only a single breeding season and suffered somewhat from small sample size. However as the only site-specific field data on piscivorous birds this information is an important contribution to the ERA.

(3.5.c) Were the estimates of exposure appropriate under the evaluation criteria, and was the refinement of analyses for the contaminants of concern (COCs) for each assessment appropriate?

Yes seems so.

(3.5.d) Were the effects metrics that were identified and used appropriate under the evaluation criteria?

These were largely literature based and no controlled toxicity studies were performed for this ERA.

(3.5.e) Were the statistical techniques used clearly described, appropriate, and properly applied for the objectives of the analysis?

Yes.

(3.5.f) Was the characterization of risk supported by the available information, and was the characterization appropriate under the evaluation criteria?

Yes, though it has to be recognized that the risk characterization is based on more limited information compared to some of the other target species. Basing risk estimates on models rather than site-specific empirical information is, and should be, more conservative and tend to overestimate the risks. However it needs to be articulated that indications of high risk from models should not be interpreted in the same way as indications of high risk from site-specific field studies. In my view the latter should trigger management actions, whereas the former should trigger collection of site-specific information to refine the risk assessment.

(3.5.g) Were the significant uncertainties in the analysis of the assessment endpoints identified and adequately addressed? If not, summarize what improvements could be made.

The uncertainties would be reduced by having additional field-based information on exposure or effects for this receptor, particularly given the large discrepancy between the modelled magnitude of risk (high) and that based on the field study (low).

(3.5.h) Was the weight of evidence analysis appropriate under the evaluation criteria? If not, how could it be improved.

This was limited due to lack of data.

(3.5.i) Were the risk estimates objectively and appropriately derived for reaches of the river where site-specific studies were not conducted?

Not performed.

(3.5.j) In the Panel members' opinion, based upon the information provided in the ERA, does the evaluation support the conclusions regarding risk to local populations of ecological receptors?

Yes.

3.6 Piscivorous Mammals

(3.6.a) Were the EPA studies and analyses performed (e.g., field studies, site-specific toxicity studies, comparison of exposure and effects) appropriate under the evaluation criteria, and based on accepted scientific practices?

The mink feeding study was appropriately conducted. Although the field surveys seemed to find very few animals, this seems to be a common feature of these types of studies.

(3.6.b) Were the GE studies and analyses performed outside of the framework of the ERA and EPA review (e.g., field studies) appropriate under the evaluation criteria, based on accepted scientific practices, and incorporated appropriately in the ERA?

The field surveys would have been strengthened by inclusion of a reference site.

(3.6.c) Were the estimates of exposure appropriate under the evaluation criteria, and was the refinement of analyses for the contaminants of concern (COCs) for each assessment appropriate?

As far as I can determine.

(3.6.d) Were the effects metrics that were identified and used appropriate under the evaluation criteria?

Yes, with the possible exception of the jaw lesions which I am not sure how to interpret in terms of impacts on mink populations. The suggestion that this effect could lead to starvation is speculation.

(3.6.e) Were the statistical techniques used clearly described, appropriate, and properly applied for the objectives of the analysis?

The statistical methods in the mink feeding study do not seem to be described in the Methods section of this study (I.3.2.1.3). They should be briefly described here.

See comments on modelling for wildlife in general.

(3.6.f) Was the characterization of risk supported by the available information, and was the characterization appropriate under the evaluation criteria?

In general yes.

(3.6.g) Were the significant uncertainties in the analysis of the assessment endpoints identified and adequately addressed? If not, summarize what improvements could be made.

Yes.

(3.6.h) Was the weight of evidence analysis appropriate under the evaluation criteria? If not, how could it be improved?

The WOE table should be separated for mink and otter because different kinds and amounts of information were available for each species.

(3.6.i) Were the risk estimates objectively and appropriately derived for reaches of the river where site-specific studies were not conducted?

Yes, these are ok.

(3.6.j) In the Panel members' opinion, based upon the information provided in the ERA, does the evaluation support the conclusions regarding risk to local populations of ecological receptors?

The ERA concluded that the risk to piscivorous mammals is high and that confidence in this conclusion is high. I believe the magnitude of the risk to piscivorous mammals is more uncertain than indicated due to the limited site-specific information on which it is based.

3.7 Omnivorous and Carnivorous Mammals

(3.7.a) Were the EPA studies and analyses performed (e.g., field studies, site-specific toxicity studies, comparison of exposure and effects) appropriate under the evaluation criteria, and based on accepted scientific practices?

Yes as far as I can determine.

(3.7.b) Were the GE studies and analyses performed outside of the framework of the ERA and EPA review (e.g., field studies) appropriate under the evaluation criteria, based on accepted scientific practices, and incorporated appropriately in the ERA?

Yes, the demographic study was a useful contribution (despite the sampling problems apparently corrected by EPA's reanalysis (or not)). GE's reanalysis of EPA's reanalysis using the same data but a slightly different statistical technique gave a non-significant result. GE treated the grids as replicates and did not weight them differentially, as a function of sample size within grids, used probit transformation, and did not include grid 3 in the analysis of males because there was only one male. In contrast, the EPA used the method of Baylor & Oris (1997; ET &C), probit transformation, and weighted grids according to the total number of organisms in the treatment. Given the dependence of the statistical significance on subtle differences between two (seemingly) appropriate statistical methods, the most robust conclusion that can be made from this study is that the response is borderline.

(3.7.c) Were the estimates of exposure appropriate under the evaluation criteria, and was the refinement of analyses for the contaminants of concern (COCs) for each assessment appropriate?

Yes.

(3.7.d) Were the effects metrics that were identified and used appropriate under the evaluation criteria?

These were based largely on literature review and used largely rat and mouse data. Though not ideal this is probably acceptable.

(3.7.e) Were the statistical techniques used clearly described, appropriate, and properly applied for the objectives of the analysis?

Yes.

(3.7.f) Was the characterization of risk supported by the available information, and was the characterization appropriate under the evaluation criteria?

This appears appropriate.

(3.7.g) Were the significant uncertainties in the analysis of the assessment endpoints identified and adequately addressed? If not, summarize what improvements could be made.

An improvement would be more site-specific information on both exposure and effects of mammals.

(3.7.h) Was the weight of evidence analysis appropriate under the evaluation criteria? If not, how could it be improved?

The WOE appears appropriate, but see comment under 3.4.h on ambiguity of WOE scoring sheets. In addition, there should be separate WOE tables for shrews and red fox due to the different kinds of information available for each (see e.g., Table 4.7).

(3.7.i) Were the risk estimates objectively and appropriately derived for reaches of the river where site-specific studies were not conducted?

No estimates of risk were made outside of the PSA.

(3.7.j) In the Panel members' opinion, based upon the information provided in the ERA, does the evaluation support the conclusions regarding risk to local populations of ecological receptors?

The ERA concluded that risks to omnivorous and carnivorous mammals is moderate to high but that there is uncertainty in these conclusions due lack of data. I agree with this conclusion.

3.8 Threatened and Endangered Species

(3.8.a) Were the EPA studies and analyses performed (e.g., field studies, site-specific toxicity studies, comparison of exposure and effects) appropriate under the evaluation criteria, and based on accepted scientific practices?

Yes as far as I can tell.

(3.8.b) Were the GE studies and analyses performed outside of the framework of the ERA and EPA review (e.g., field studies) appropriate under the evaluation criteria, based on accepted scientific practices, and incorporated appropriately in the ERA?

None performed.

(3.8.c) Were the estimates of exposure appropriate under the evaluation criteria, and was the refinement of analyses for the contaminants of concern (COCs) for each assessment appropriate?

Yes they seem so.

(3.8.d) Were the effects metrics that were identified and used appropriate under the evaluation criteria?

There seem to be very important uncertainties here, given the lack of relevant effects data for T&E species.

(3.8.e) Were the statistical techniques used clearly described, appropriate, and properly applied for the objectives of the analysis?

No additional comments.

(3.8.f) Was the characterization of risk supported by the available information, and was the characterization appropriate under the evaluation criteria?

If it can be assumed that T&E species are such, not because they are particularly sensitive to PCBs and other toxicants, but rather because they have life-history characteristics that make them sensitive to demographic impacts, then it is appropriate to use effects thresholds from non T&E species in the assessment as was done here. However it should be recognized that the same impairment of survival or reproduction could have much greater consequences for T&E populations than for non T&E populations.

There was a problem with the risk estimates differing for different life stages of the bald eagle and an average taken as an estimate of overall risk. This is inappropriate as discussed under Charge Question 2. A life cycle model could have been applied to calculate elasticities (sensu Caswell) of the different life stages to quantify their relative importance to population dynamics of this species.

(3.8.g) Were the significant uncertainties in the analysis of the assessment endpoints identified and adequately addressed? If not, summarize what improvements could be made.

Yes, and although they are large, I cannot see how they can be reduced further without collecting additional data.

(3.8.h) Was the weight of evidence analysis appropriate under the evaluation criteria? If not, how could it be improved?

It is extremely limited due to lack of data.

(3.8.i) Were the risk estimates objectively and appropriately derived for reaches of the river where site-specific studies were not conducted?

Appears ok.

(3.8.j) In the Panel members' opinion, based upon the information provided in the ERA, does the evaluation support the conclusions regarding risk to local populations of ecological receptors?

The ERA concludes that the risk is high for selected threatened and endangered species but that there is uncertainty in this conclusion. I would rate the risk as very uncertain given the heavy emphasis on modelling results (which for all wildlife gave higher risk estimates than field or site-specific data where these were available).

Charge Question 4. Are the summary discussions and conclusions in the ERA supported by the information provided in the report, and did the conclusions describe the risks in an objective, reasonable, and appropriate manner?

General Comment on WOE - In their ecological risk assessment guidelines, EPA (1998) states: "A major advantage of field surveys is that they provide a reality check on other risk estimates, since field surveys are usually more representative of both exposures and effects (including secondary effects) found in natural systems than are estimates generated from laboratory studies or theoretical models." Thus it should not be particularly surprising that field surveys do not always match the risks predicted from laboratory studies or models. The above statement would support weighting field studies higher than either lab or model outputs, when these provide conflicting estimates of risk. Although field studies also have their limitations these need to be carefully judged in comparison with the limitations of lab/model extrapolations. I am not sure that this has been consistently and objectively done in the present ERA. Given the subjectivity in the qualitative WOE used here, I recommend that the relative weightings of laboratory, literature, model, and field results be given further review.

The HQ analyses should be modified to consider the size of the HQ (not simply is it greater than one?) and some measure of probability or proportion of samples exceeding the critical HQ of 1.

In my view the issue of risk acceptability should not be addressed in the ERA. The ERA should focus on quantifying risks and their associated uncertainties. Acceptability is a risk management issue.

I would recommend that use of the term 'significant' (e.g., as in significant risk) be restricted to situations in which reference is being made to statistical tests.

Charge Question 5. To the best of the Panel's knowledge, is there other pertinent information available that was not considered in the ERA? Is so, identify the studies or data that could have been considered, the relevance of such studies or data, and how they could have been used in the ERA.

The amount of information considered in this ERA is impressive. I have no knowledge of additional studies or data that could have been considered. Indeed, the copious amounts of

data and analyses made reviewing the ERA in a realistic time frame rather challenging. Providing a good roadmap (i.e., Fig 1.1-3) is thus essential. Additional guidance on finding relevant pieces of information would be helpful.

Regarding the overall structure of the ERA, I can appreciate the value of placing the technical details in Appendices for the particularly interested reader. However, I suspect that most readers of this document will either be satisfied with the Executive Summary (perhaps slightly expanded) or will want all of the details. Thus in my view the main chapters provide too much information for the casual reader and too little for the expert.

There were several places in the ERA in which documents 'in preparation' were cited. This should be avoided as a matter of good practice.

In the limited time that I have been involved in this project it has been my impression that the cooperation and collaboration between the GE team and the EPA team have been exceptionally positive and constructive. This, I suspect, has not only made the process more pleasant, but I believe has led to a better risk assessment than would have otherwise been done. This ultimately benefits the Housatonic and all of the stakeholders with interest in it. If it is feasible to formally share the lessons learned in a form that could provide guidance to other risk assessments of this nature, I would strongly encourage that this been done.