6.0 ANALYSIS AND ASSESSMENT OF SAMPLING DATA

Sampling data will be analyzed with analysis results interpreted and evaluated consistent with the purpose, goals, and objectives of the BEMP to assess Basin environmental conditions and trends and document progress toward and attainment of the ecological benchmarks identified in the ROD. As discussed in preceding sections, sampling data represent measurements at selected monitoring locations and times of monitoring parameters that include chemical concentrations in surface waters, sediments, and biota; chemical loading and AWQC ratios in surface water; and other ecologically relevant parameters. In particular, Section 4 identified by media the monitoring parameters and sampling schedules.

This section provides a general discussion of the analysis and assessment of the sampling data. Section 6.1 covers data analysis, with an emphasis on statistical hypothesis testing consistent with the discussion of Section 3. Section 6.2 discusses the follow on assessment (interpretation and evaluation) of the data analysis results. Assessment will be framed within the purpose, goals, and objectives of the BEMP using an adaptive management strategy that supports the 5-year remedy reviews required by CERCLA.

The BEMP assumes that extensive analysis of accumulated monitoring data will be conducted at 5-year intervals timed to precede the 5-year remedy reviews required by CERCLA. These 5-year data analyses will follow the approach discussed in Section 6.1 and also include the assessment of results discussed in Section 6.2. Analyses and assessments will be documented in BEMP Technical Memoranda, which will be used to support the 5-year remedy reviews.

Limited-scope data summary reports will be issued yearly. The yearly data summary reports will include tabular and graphical summaries of the monitoring data, with analysis limited to computation of standard sample statistics. The yearly reports will identify any potentially significant "anomalies" that may require early attention.

Also, it is anticipated that as they become available, monitoring data will be accessible on the web for inspection by the interested public. Data management is discussed in Section 7.

6.1 DATA ANALYSIS

This section provides a conceptual basis for a general understanding of how sampling data resulting from the sampling identified in Section 4 will be analyzed to statistically test the hypotheses discussed in Section 3. Supplementary and complementary analyses for probabilistic characterization of monitoring parameters – including confidence levels and intervals, probability distributions, and statistical power analyses – are also discussed here. While the discussions generally apply to all the media and monitoring parameters identified in

Section 4, application to specific media may evolve as needed for each medium over the implementation period of the BEMP. The discussions are thus intended as adaptable guidelines.

The sample data represent time-specific measurements of monitoring parameters at Basin sampling locations for surface water, sediments, and biota. Because monitoring parameters reflect naturally variable temporal and spatial averages, the sampling data are analyzed as time-varying aggregate averages applicable to their specific monitoring locations.

The focus is on statistically analyzing the sampling data in a framework that considers the hypothesis testing discussed in Section 3 under conditions of uncertainty. The sampling data are analyzed as aggregated averages and associated time-history trends. The conceptual overview of this section is extended in the forthcoming BEMP Technical Memorandum to include a more detailed technical discussion and quantitative development, as described in Appendix D.

The statistical analyses deal with the uncertainty inherent in the sampling data in a scientifically defensible manner. Yet scientifically defensible means neither perfect nor uniquely objective. The statistical analyses, like all analyses, require professional judgment, and results must be fairly interpreted in proper context(s) using professional knowledge and insight. It is also expected that the analyses will be supplemented and complemented with applicable information available from other sources, including the results from remedy effectiveness monitoring. Interpretation and evaluation of the statistical analysis results are discussed in Section 6.2.

6.1.1 Natural Variability, Uncertainty and Statistical Analyses

The sampling data will be limited in number and accuracy and subject to inherent natural variability and statistical fluctuations—common effects in all complex natural systems like the Basin. Coupled with the uncertainty of natural variability is the statistical uncertainty of limited sampling measurements having imperfect accuracy (i.e., random measurement "error"). The net effect is that exact true values of monitoring parameters cannot be known with certainty, but must be estimated from statistical analysis of the available sampling data.

Because true values are uncertain, measuring progress toward benchmarks and improvements in environmental conditions requires statistical analyses of the sampling data. Statistical analysis quantitatively characterizes the uncertainty in the true values reflected in the sampling data. Statistical estimates thus represent *potential* true values inherent in the sampling data. The statistical analyses characterize the sampling data in terms of probabilities and associated terminology of hypothesis testing, as discussed next.

6.1.2 Measuring Remedy Progress and Attainment of Benchmarks

Measurable progress toward ecological improvements and benchmarks means that there are acceptable *probabilities* or "confidence levels" that the true values of measured monitoring parameters have generally improved over time. Systematic changes over time are represented by a time-history trend (trend) at a given monitoring location. It may be concluded that there is a true trend if the sampling data for the given monitoring location indicate a trend at an acceptable confidence level.

Concluding that a quantified benchmark associated with a given monitoring parameter (e.g., AWQC ratio) has been met uses the same approach as used for trends: there must be an acceptable confidence level or probability that the inherently uncertain true value (e.g., AWQC ratio) has actually met the benchmark. It may be concluded that the benchmark has been met if the sampling data meet the benchmark at an acceptable confidence level.

The BEMP is designed to support quantification of both the trends over time of the monitoring parameters and the probability (confidence) that the parameters meet applicable benchmarks. Data from the monitoring program will be analyzed using common statistical techniques to estimate the true values and trend over time.

Measuring remedy progress and attainment of benchmarks is based in large part on statistical hypothesis testing discussed in Section 3. The burden of proof used to test the hypotheses is quantified by the an acceptable confidence level, approximated as "1–alpha" where alpha is the "significance level." Numerical examples of confidence levels include 0.95, 0.90, or 0.51 with corresponding alphas of 0.05, 0.10, and 0.49. Effective hypothesis testing also includes estimating "statistical power," which measures the ability of the sampling data to detect specified magnitudes of change in the monitoring parameters. For a given parameter, location, and sample design, statistical power increases with sample size, the number of independent measurements in the sample, and symbolized as "N." Hypothesis testing, including confidence levels and statistical power, will be discussed further in the following sections.

The hypothesis testing will be supplemented and complemented by estimating statistical confidence intervals and limits on the true parameter averages and true slopes of the trend lines, based on the sampling data. This approach is consistent with standard scientific and statistical principles.

6.1.3 Hypotheses Testing and Decision Criteria

Statistical testing will compare each monitoring hypothesis (the alternative hypothesis) against its corresponding null hypothesis. The null hypothesis is a presumption that is accepted (but not "proven") unless statistically "falsified," and hence rejected in favor of adopting the alternative hypothesis. This approach places the burden of proof on rejecting the null hypothesis or, equivalently, on accepting the alternative hypothesis.

As in any complex natural system like the Basin, natural variability and statistical fluctuations in monitoring parameters (and sampling limitations, including random measurement error) means that statistical hypothesis testing suffers from potential error. Error rates or probabilities associated with hypothesis testing are characterized as Type I and Type II errors, measured by alpha and "beta."

- Rejecting a true null hypothesis is a Type I error, measured by alpha, the significance level of the statistical test. Recall that the complement of alpha (1 – alpha) is the confidence level of the test. Assume, as a simple example, that there is *in fact* no zinc reduction (zero change) at a particular monitoring location and that statistical testing used an alpha of 0.05. In this case, falsely rejecting a null hypothesis that there is no zinc reduction would be expected to occur on average in 5 out of 100 measurements, or in 5 percent of repeated measurements (alpha = 0.05). Correctly accepting the null hypothesis would be expected in 95 percent of repeated measurements, on average (1 – alpha = 0.95).
- Accepting a false null hypothesis is a Type II error, measured by beta, the probability of accepting a false negative hypothesis. The complement of beta (1 – beta) is the statistical power of the test: the ability to detect a specified magnitude of change in the monitoring parameters. Assume, as a simple example, that there is *in fact* a zinc reduction (of a certain magnitude) at a particular monitoring location and that statistical testing used a beta of 0.20 (for that magnitude). In this example, falsely accepting a null hypothesis that there is no zinc reduction would be expected to occur on average in 20 out of 100 measurements, or in 20 percent of repeated measurements (beta = 0.20). Correctly rejecting the null hypothesis would be expected in 80 percent of repeated measurements, on average (1 – beta = 0.80).

Equivalently, erroneously accepting a false alternative hypothesis is a Type I error and erroneously rejecting a true alternative hypothesis is a Type II error. Type I and Type II errors (i.e., alpha and beta) are related by a statistical power analysis (see Appendix D). These potential statistical testing errors are consistent with the scientific fact that hypotheses cannot be "proven" true or false, except in the inductive statistical sense—by the weight of evidence inherent in the representative data.

6.1.3.1 *Null Hypotheses*

It is important to be clear that null hypotheses are presumptions. In more precise statistical terms, accepting a null hypothesis means "failing to reject" that null hypothesis. Accepting a null hypothesis is thus not a test of truth. A false null hypothesis may go undetected because of limited data having inadequate statistical power to reject, at an acceptable confidence level, false null hypothesis. This type II error (beta) illustrates why, ideally, there should be adequate statistical power to "test" the validity of the null hypothesis.

To reiterate, accepting a null hypothesis does not statistically "prove" or validate that null hypothesis. Accepting a null hypothesis absent compelling contrary information is a policy decision. A false null hypothesis may thus be accepted by presumptive default. This argues for why null hypotheses should represent a protective policy position, a conservative position (often the status quo) that if false results, on balance, in less aggregate expected cost (loss and risk) than if it were true but assumed false.

The null hypothesis acts as the hurdle or burden of proof that must be met by the available monitoring data to accept the alternative hypothesis. If the available data fails the burden-of-proof test, thus failing to reject the null hypothesis, the null hypothesis -- whether true or false -- is accepted. If the available data clears the burden-of-proof test, the null hypothesis is considered false and the alternative hypothesis is accepted as true.

The status quo for active Superfund sites represents conditions that are not protective and thus require cleanup to effect a change to being protective. Null hypotheses for Superfund sites are, therefore, typically formulated as "not-protective" or, equivalently, "no-change" from the status quo. The not-protective or no-change presumption stands until "proven" false by the data. Setting the null hypotheses as "not-protective" or "no-change" is a conservative position from the point of view of environmental protection.

This Superfund approach is used in the BEMP, where the null hypotheses represent "no-change" and the alternative hypotheses represent positive "change" in terms of improving conditions toward cleanup. As detailed in Section 3, the monitoring hypotheses are alternative hypotheses (positive change), evaluated against the null hypothesis of no change. Hypothesis testing for potentially degrading conditions (negative change) is discussed further in Section 6.1.4.

6.1.3.2 *Choosing Decision Criteria*

Clearly, the hypothesis testing requires specifying appropriate decision criteria for acceptable Type I and Type II error probabilities, as represented by alpha and beta. *Selecting values for alpha and beta are subjective risk management decisions*, as there are no uniquely correct values. Ideally, alpha and beta appropriately reflect the risk and cost associated with potential decision errors. The following discussion provides additional background for selecting alphas and betas.

- A maximum acceptable Type I error probability, or alpha, represents the acceptable burden of proof to reject the null hypothesis, and thus accept the alternative hypothesis. The burden of proof increases with decreasing alpha. Quantity 1–alpha, the confidence level associated with rejecting the null hypothesis, increases as alpha decreases. The confidence level increases with the burden of proof.
- Admissible Type I errors cannot exceed 0.50, which means that admissible alpha cannot exceed 0.50. Alpha equal to 0.50 is a limiting condition corresponding to no null hypothesis. Practically, the maximum admissible alpha is 0.49 (theoretically 0.4999…). Alpha greater than 0.50 is equivalent to exchanging the null and alternative hypotheses, effectively bringing alpha back to less than 0.50.

Alpha should be situation-dependent, and rationally consistent with the expected cost of falsely rejecting a true null hypothesis. Following historic convention, alpha has commonly been set at 0.05 in scientific studies and EPA's CERCLA regulatory-guidance documents. Notably, however, alpha values between 0.20 and 0.05 are recommended in EPA's recent *Guidance for Comparing Background and Chemical Concentration in Soil for CERCLA Sites* (EPA 540-R-01-003, September 2002). It remains useful to recognize that alpha may, in noncritical cases, be as high as 0.49, which represents a burden of proof corresponding to "more probable than not." While low alpha values are appropriate for helping assure adequate protectiveness and conservative decision-making, the BEMP reserves the flexibility to use higher alpha values, if appropriate, for what are essentially non-decision situations. Again, there is no single correct value of alpha applicable in all situations. After a short discussion of beta, the concluding paragraphs of this section introduce a tiered-approach to choosing alpha and beta in the context of adaptive management.

An acceptable beta depends on the required or desired statistical power, 1–beta, to detect a true alternative hypothesis, and thus reject the (false) null hypothesis. Whereas alpha may be a single-valued decision criterion, beta and statistical power are more complex. Statistical power increases with increasing alpha. Power also depends on the variability of the data, the sampling design and sample size, and the minimum "effect size" – the magnitude of the effect – to be detected. The effect size is the magnitude of the difference between the null hypothesis and the true value. Statistical power is estimated by analysis, as detailed in Appendix D.

As with alpha, there is no single correct value of beta or power. As with alpha, beta should be situation-dependent and consistent with the expected cost of failing to accept a true alternative hypothesis.

While uniquely correct values do not exist, alpha and beta, as risk management decision criteria, should be realistic and balanced to minimize expected costs (risks). A general aim is to maximize cost-effectiveness while maintaining acceptable protectiveness. Appropriate values are therefore likely to vary over time and between monitoring parameters and media. Particularly in the context of adaptive management, a "tiered approach" related to the severity of real or potential decision consequences (expected costs) may be useful for choosing alpha and beta values:

- A low-consequence tier with alpha potentially as high as 0.49 could be used for testing trends of monitoring parameters that do no relate to a specific quantitative benchmark or result in significant costs of actions or inaction. Similarly, low power could be acceptable where uncovering a false null hypothesis has minor consequences—e.g., where maintaining the status quo, consistent with maintaining the null hypothesis, is not costly.
- Conversely, a high-consequence tier with alpha of 0.05 would be used where the costs of falsely rejecting a null hypothesis are considered high—e.g., declaring that a

quantitative benchmark has been achieved where it has not. High power could be needed to uncover a false null hypothesis that maintains a costly status quo or blocks acceptance of improved understanding or management practices.

Although specifics have not been developed at this time, the BEMP assumes that a tiered approach to choosing alpha and beta will be used in the adaptive management framework outlined in Section 6.2. Thus, for a given monitoring parameter (e.g., zinc concentrations) there may be multiple tiers of alpha and beta for different effect magnitudes of interest, which may evolve over the course of implementing the remedy, including the BEMP. Clearly, the real constraints of available funding must be considered, particularly where large number of samples would be needed for small values of alpha and beta.

A tiered approach is consistent both with principles of adaptive management and EPA's "ideal approach to hypothesis testing," as stated in their data quality assessment guidance (EPA 2000, p5-12,13). EPA 2000 *Guidance for Data Quality Assessment*, QA/G-9, Final July 2000. Characteristics of EPA's ideal approach include the following. It sets up the null hypothesis to protect the environment. It controls the false rejection error (alpha). It encourages quality in term of high precision and accuracy, and thus statistical power. Yet the ideal also seeks to minimize expenditures in situations where decisions are relatively easy—e.g., all measurement observations are far from decision thresholds or levels of serious interest.

6.1.4 Hypothesis Testing for Potentially Degrading Conditions

Recall from Section 3 that the monitoring hypotheses, which represent alternative hypotheses, have been generally formulated as improvements in monitoring parameters. The resulting hypothesis tests are thus "improvement vs. no-improvement," consistent with the remedy intent to improve ecological conditions.

Detecting potentially degrading conditions is also important. Although the null hypotheses explicitly represent no-change (as no-improvement), they *implicitly* include degradation in the monitoring parameters being analyzed. However, to explicitly account for potentially degrading conditions, complementary hypotheses of the form "degradation vs. no-degradation" will also be tested. These complementary 1-tailed tests are considered superior to single 2-tailed tests (using a significance level of ½ alpha) because the direction of the change (improvement or degradation) is explicit. In a classical 2-tailed test, direction is not explicit.

6.1.5 Limitations of Hypothesis Testing

By itself, classical hypothesis testing is an inadequate basis for interpreting actual conditions from sampling results of monitoring programs. In particular, significance or confidence levels may suggest a simple "yes-no" answer to the validity of monitoring hypotheses. Without explicit consideration of the probabilities associated with the range of potential values of the true (but uncertain) monitoring parameter, this "yes-no" interpretation can lead to apparent dilemmas.

For example, a monitoring hypothesis may be rejected at significance level alpha of 5% (95% confidence level) yet accepted at alpha of 10% (90% confidence level). This example illustrates that the choice of alpha, or confidence level, is not a fundamentally scientific issue. Rather, the choice of alpha is always a risk management decision, determined by policy or cost-consequence evaluation. Decisions will vary with circumstances and contexts, which will generally vary over time.

Therefore, as part of data analysis and interpretation, hypothesis testing will be supplemented with explicit estimates of the uncertainty in the monitoring parameters, as characterized by probability curves. This approach leads to a more general form of hypothesis testing related to confidence intervals and limits associated with the true values.

6.1.6 Confidence Intervals and Limits

Confidence intervals and limits will be estimated to provide a more complete characterization of the sampling data and their implications to actual conditions in the Basin. Confidence intervals and limits will be estimated both for the true average of each monitoring parameter for each sampling event and the associated trend over time. Estimates will be used to quantify the uncertainty in the true values and support a generalized approach to hypothesis testing. Uncertainty is quantified as cumulative probabilities.

6.1.7 Cumulative Probability Curve Application to Hypothesis Testing

Cumulative probability curves can be used to determine the maximum alpha level, or minimum confidence level, at which a given null hypothesis would be rejected. The maximum alpha level is the "critical alpha"; the minimum confidence level is the "critical confidence level." For given sampling data, the critical alpha is the maximum alpha that would result in rejecting the null hypothesis. The critical confidence level is the minimum confidence level, consistent with the critical alpha.

Whether formally established or hypothetical, the given null hypothesis being evaluated for a critical alpha can be any potential *value of interest* of the true value (average or trend) of the monitoring parameter. The values of interest may be any "target value," or benchmark, including ROD ecological benchmarks.

The monitoring hypotheses of Section 3 use a null hypothesis of no (zero) trend; in these cases, the target value is zero. More comprehensive analysis and interpretation of BEMP sampling data will assess ranges of possible target values, and thus use corresponding ranges of hypothetical null hypotheses.

This generalized approach to hypothesis testing allows entire ranges of target values (null hypotheses) and potential true values to be analyzed in a systematic, practical and rapid way. Results can be assessed against benchmarks in terms of sensitivity to target values and critical alpha values. This capability may be particularly useful to risk management decision making and adaptive management.

6.1.8 Post-Sampling Statistical Power Analyses

During development of the BEMP, pre-sampling statistical power analyses were used to analyze the effectiveness and efficiency of proposed and alternative sampling designs for hypothesis testing and characterizing uncertainty in true values. These analyses and associated supporting discussions are included in Appendix D. Analysis results were evaluated during development of the BEMP within the overall context determined by the ROD, basin conditions, the data quality objectives (DQO) process, and various practical tradeoffs and limitations, including constraints imposed by the projected availability of funds. The BEMP sampling designs identified in Section 4 reflect the result of this overall evaluation.

The pre-sampling power analyses formally considered currently available quantitative baseline data for surface water and sediments, as summarized in Section 2.3. However, because statistical analyses of baseline biological data were not available, formal power analyses were not specifically conducted for biological-sampling designs. Biological sampling designs were developed using professional judgment and suspected acceptable power, subject to practical constraints, and interpretation of power analysis results for surface water and sediments.

For all media (surface water, sediments, and biological), power analyses will be used in the post-sampling analysis of the actual, realized sampling data that will be obtained as the BEMP is implemented. Post-sampling analyses are similar to pre-sampling analyses except actual (post) sampling data is analyzed in the former. Post-sampling power analyses allow quantitative evaluation of the evidential support provided by the sampling data for both null hypotheses and monitoring (alternative) hypotheses.

Post-sampling power analyses also estimate probabilities of accepting or rejecting null hypotheses (target values) and alternative hypotheses conditional on assumed true values. While true values are always uncertain to the extent there is natural variability and sampling limitations, hypothetical true values may be *assumed* and statistically analyzed for implications. These analyses use the actual sampling design and realized data, which reflect the extent uncertainty affecting estimates. Post-sampling power analyses are thus "what if" analyses that complement the generalized approach to hypothesis testing discussed in the previous section.

The post-sampling power analyses, conducted after the sampling data are available, provide the most accurate estimates of the actual sampling design "performance." Since post-sampling analysis results are not available now, the adaptive management strategy allows appropriate modification of the BEMP (within practical constraints) should statistical power prove inadequate in the future.

6.1.9 Methods of Analysis

The BEMP will appropriately use or adapt common graphical and mathematical statistical methods appropriate for exploring, characterizing, or analyzing data from observational studies, consistent with the discussions in this document. These methods generally include time-history graphs, summary statistics, distribution analyses, various quantitative statistical tests, power analyses, and regression methods. Appendix D and the forthcoming BEMP Technical Memorandum provide a more detailed background discussion including pertinent mathematical details. The methods assumed in this section are partially predicated on results and experience gained from analyzing currently available Basin environmental (background) data in the context of conducting the RI/FS and developing the ROD, as well as supplementary exploratory analyses done in support of the BEMP.

The BEMP recognizes that results from statistical analyses of complex, real-world monitoring data represent probabilistic estimates of uncertain true field values and that explicit probability levels represent approximate statements of knowledge that are always conditional on available information and its interpretation. That theoretical statistical models are never perfectly consistent with real world data is also recognized. Furthermore, the BEMP data may be statistically "messy" because of the dynamic and complex nature of the Basin along with the practicalities and contingencies associated with multiple sampling and analysis programs conducted over many years.

In general, the selection and use of statistical methods should adequately consider the tradeoffs between method benefits and limitations. Potential limitations include effects of deviations between the observational data and model assumptions (e.g., data or parameter correlations, potentially mixed populations or outliers, distribution forms of parameters or populations, gaps in the sampling record, and so on). While there are no hard and fast rules, in certain cases, multiple analysis methods and interpretations may be appropriate. It is also likely that data analysis needs will evolve over time.

In practical terms, these considerations argue for the BEMP assuming a flexible viewpoint that maintains appropriate scientific rigor and rationality while avoiding unnecessary restrictions or complications imposed by rigid interpretations of academic fine-points or unobtainable expectations of objectivity or precision. In support of this practical viewpoint, generalized Bayesian concepts (including weight-of-evidence arguments using formal or informal information "updating") may be used. It bears emphasis that appropriate analysis, interpretation and evaluation requires sound professional judgment exercised in the context of the purpose, goals, and objectives of the BEMP, with adequate understanding of relevant Basin processes. Formalized methods and criteria cannot eliminate the fundamental need for professional judgment exercised in real time.

Clearly, because the BEMP is expected to evolve over its 30-year implementation period, this current version of the BEMP does not limit future beneficial use of other applicable methods that may become appropriate during implementation. The concepts and methods discussed here may be appropriately updated and modified as evolving conditions may dictate. This evolutionary approach is consistent with the adaptive management strategy inherent in the BEMP, as discussed in the next section.

6.2 ASSESSMENT AND ADAPTIVE MANAGEMENT

Following the analysis of sampling data generally discussed in Section 6.1, the analysis results will be interpreted and evaluated consistent with the relevant context discussed in Sections 1 through 3 and the purpose of the BEMP to periodically monitor, quantify, and document overall remedy performance. As discussed in Section 3, remedy performance is defined in terms of meeting or progressing toward the ecological benchmarks identified in the ROD. Those benchmarks represent the ecological "performance expectations," the goals and objectives, of the remedy that will be measured and assessed by the BEMP.

The integrative assessment of data analysis, interpretation, and evaluation provides the information needed for periodic evaluation of evolving remedy performance against the ROD benchmarks, the expectations of remedy performance. Potentially, these evaluations may help identify and guide "course corrections" to remedy implementation that improve remedy performance, including cost-effectiveness. Specific efforts include detecting trends or major trend discontinuities, which may signal a need to update critical assumptions or change management practices, potentially including the BEMP or the remedy itself.

The integrated assessment used in the BEMP will be conducted within an adaptive management framework, as called for in the ROD. In general terms, adaptive management is a systematic strategy for continually learning from the ongoing monitoring results to cost-effectively improve future remediation and monitoring. It provides a purposeful feedback loop to assess evolving conditions and identify useful changes to the remedy, including long-term monitoring, as identified in the BEMP. Adaptive management is a key strategic component inherent in the BEMP.

6.2.1 Adaptive Management Framework

Assessment of the evolving Basin ecological conditions against the expectations of remedy performance, as discussed in preceding sections, provides a basis for responsive adaptive management. The adaptive management framework will be integrated with the CERCLA five-year review process to provide the regulatory basis for implementing practical adaptive management changes that may be appropriate to improve the remedy, including the BEMP, as it is being implemented. Because there are numerous site-specific remedies planned under the ROD over the next 30 years, the adaptive management framework will consider information gathered from the BEMP and the effectiveness monitoring programs, as well as relevant information available from the LEMP (Lake Environmental Monitoring Program) or any other pertinent monitoring. The following 3-step assessment strategy will be implemented as an adaptive management framework:

1. Analyze BEMP sampling data consistent with Sections 3 through 6 and supporting appendices. Consider, as appropriate, other applicable and available data, including effectiveness monitoring and LEMP results.

2. Interpret and evaluate results, update understanding of Basin ecological conditions, and compare results against ROD benchmarks and the status of remedy implementation.

3. Decide what improvements, if any, are appropriate to modify the BEMP, effectiveness monitoring, or remedy design and implementation.

Within the integrated 3-step strategy, the following related questions will be answered. First and foremost, "is the remedy functioning as intended by the ROD?" Focused answers will provide interpretation and evaluation of the results of the statistical analyses of the cumulative monitoring data in terms of the monitoring hypotheses identified in Section 3. The intent is to document the magnitude and geographic extent of remedy performance in terms of the ROD benchmarks. To reiterate from Section 3, ecological benchmarks include:

- Decreases in dissolved zinc and cadmium concentrations in surface water.
- Decreases in particulate lead concentrations in the flood plain soils/sediment, levees, and riverbed sediments.
- Decreases in particulate lead loads and concentrations in surface water.
- Decreases in zinc AWQC ratios
- Improvements in biotic benchmarks
- Improvements in metals retention in CDA Lake
- Identification of any "unwanted" impacts from remedy implementation
- Clear indications of progress toward achieving benchmarks

As discussed in Section 6.1, statistical analyses will include quantification of trends over time of monitoring parameters and the probability (confidence) that the parameters meet applicable benchmarks. Quantitative results will be complemented and supplemented by narrative that considers the broader context of Basin conditions and the state (relevant history) of remedy implementation at the time of the assessment.

A second question that will be answered is "does interpretation and evaluation of available data from the BEMP and other monitoring programs (LEMP, site-specific remedial actions, other pertinent programs or data sources) suggest new or refined understanding of Basin processes that are relevant to the remedy?" Focused answers to this broad question will be provided in narrative form that is integrated with answers to the question of remedy function, as discussed in the preceding paragraph.

Answers to a third set of questions will focus on identification of any warranted revisions or modifications to the BEMP or site-specific remedial action monitoring. The intent is to improve the technical performance or cost-effectiveness of the monitoring, including changes that may be needed to meet budget or other practical considerations. Related questions aimed at the BEMP, which will be answered in narrative form, include:

- Which monitoring efforts (if any) can be reduced or eliminated?
- Are there monitoring elements that should be added?
- Are the monitoring stations still appropriate, considering remedial action plans and other factors?
- Are there new monitoring techniques that should be considered?
- What changes (if any) should be considered to the statistical analysis techniques used to measure and identify progress?

Although not explicitly part of the BEMP, other adaptive management questions that will be addressed as part of the five-year review include (but are not limited to) the following. Are the exposure assumptions, toxicity data, cleanup levels, and remedial action objectives used at the time of remedy selection still valid? Has any other information come to light that could call into question the protectiveness of the remedy?

6.2.2 Reporting of BEMP Results

Major assessments that include the Section 6.1 analyses with interpretation and evaluation of results using the adaptive management framework will be conducted at 5-year intervals timed to support the 5-year remedy reviews required by CERCLA. Documentation of these efforts will be formalized in BEMP Technical Memoranda.

For years where BEMP Technical Memoranda are not developed, data summary reports will be issued. The yearly reports will include tabular and graphical summaries of the monitoring data, including updated time histories. Analysis in the yearly reports will be limited to computation of standard sample statistics. Interpretation and evaluation in the yearly reports will be limited to identification of any potentially significant "anomalies" or concerns that may require early attention, before consideration in the 5-year reports.