# APPENDIX L

## STATISTICAL METHODOLOGY FOR BENTHIC INVERTEBRATE COMMUNITY ECOLOGICAL ASSESSMENT ENDPOINT ANALYSES

# Appendix L: Statistical Methodology for Benthic Invertebrate Community Ecological Assessment Endpoint Analyses

## Final Aquatic Baseline Ecological Risk Assessment—Upper Columbia River

PREPARED FOR
Teck American Incorporated

DATE
July 2026

REFERENCE
0812334



# CONTENTS

TABLES

FIGURES

## LIST OF TABLES



## LIST OF FIGURES



## LIST OF EQUATIONS

## ACRONYMS AND ABBREVIATIONS

| | |
|---|---|
| AIC | Akaike Information Criterion |
| AOI | area of interest |
| BERA | baseline ecological risk assessment |
| BMI | benthic macroinvertebrate |
| D | Cook's distance |
| EC | effects concentration |
| EC20 | effects concentrations at which a 20 percent reduction in the biological response relative to the upper asymptote is observed |
| EC50 | effects concentration at which a 50 percent reduction in the biological response relative to the upper asymptote is observed |
| EPA | U.S. Environmental Protection Agency |
| EPT | Ephemeroptera, Plecoptera, and Trichoptera |
| IV | indicator value |
| LL3 | three-parameter sigmoidal model |
| LM | linear model |
| LOE | line of evidence |
| LOESS | locally weighted polynomial regression |
| n | sample size |
| NMDS | nonmetric multidimensional scaling |
| NS | natural spline |
| OU | Operable Unit |
| PC | principal component |
| PCA | principal components analysis |
| $R^2$ | coefficient of determination |
| Site | Upper Columbia River site |
| SD | standard deviation |
| TOC | total organic carbon |
| UCR | Upper Columbia River |

# 1. INTRODUCTION

This appendix describes the statistical methods used to complete data analyses associated with the assessment of the benthic invertebrate community endpoint in the Aquatic Baseline Ecological Risk Assessment (BERA) of the Upper Columbia River site (UCR; hereafter, the Site[1]), which extends from the U.S.-Canada border (River Mile 745) to Grand Coulee Dam (River Mile 596). The analyses described herein were used in both the exploratory phase of the data analysis as well as the analyses used to address the lines of evidence (LOEs) for the benthic invertebrate community assessment endpoint. Detailed findings of these analyses using the statistical methods described herein are provided in the main body of the Aquatic BERA and associated appendices. For ease of reference, this appendix is organized by LOE for the benthic invertebrate community ecological assessment endpoint.

The specific analyses described herein include:

- Section 2: General data visualizations used for many of the benthic invertebrate community LOEs;
- Section 3: Exploratory and contextual analyses including principal components analysis (PCA) and nonmetric multidimensional scaling (NMDS) of benthic invertebrate community data (LOE 3);
- Section 4: Bivariate and multivariate modeling used in the concentration-response modeling for the bioassay and chemistry LOE (LOE 4);
- Section 5: Bivariate modeling methodologies for the benthic invertebrate community metric and chemistry LOE (LOE 5); and
- Section 6: Reference envelope and central tendency tests for the bioassay and benthic invertebrate community metric LOEs (LOE 2 and LOE 3).

This appendix is intended for a technical audience with statistical expertise and that is interested in additional specifics around the statistical methodology used to complete the LOE analyses for the benthic invertebrate community ecological assessment endpoint. Justification for the statistical approaches and the specifics of how they were implemented, including the data used to execute the analyses, is included in the main body of the Aquatic BERA.

[1] The Site as defined within the June 2, 2006, Settlement Agreement is the areal extent of hazardous substances contamination within the U.S. in or adjacent to the UCR, including the Franklin D. Roosevelt Lake, from the U.S.-Canada border to Grand Coulee Dam, and those areas in proximity to the contamination that are suitable and necessary for implementation of response actions (USEPA 2006).

# 2. DATA VISUALIZATIONS

Several data visualization approaches are used throughout the assessment of benthic invertebrate communities in the Aquatic BERA. Boxplots are used to illustrate group comparisons, scatterplots are used to illustrate correlations between two variables, and mosaic plots are used to distill and present results from analyses of numerous variables. This section provides guidance on how to interpret the contents of each of these commonly used graphical approaches in the Aquatic BERA.

## 2.1 BOXPLOTS

Boxplots are used throughout the benthic invertebrate community assessment sections of the Aquatic BERA to illustrate group comparisons such as how bulk sediment concentrations of metals compare among areas of the Site. Boxplots were produced using the `geom_boxplot()` function in the `ggplot2` package in R (Wickham et al. 2019); all boxplots present the first and third quartiles (25th and 75th percentiles, respectively) as the lower and upper boundaries of the box (also known as the "hinges"). The horizontal line through the box represents the median value (50th percentile), and whiskers (i.e., vertical lines) extending out of the top and bottom of the box represent maximum values no farther than 1.5 times the interquartile range. For boxplots displayed on a log-scale y-axis, the boxplot statistics are calculated on log transformed data and then back transformed for visualization. In these cases, the whiskers are calculated as follows (Equations 1 and 2):

### EQUATION 1: UPPER WHISKER FOR BOXPLOTS

$$Whisker_{upper} = e^{\ln(Q3)+1.5*\ln(\frac{Q3}{Q1})}$$

### EQUATION 2: LOWER WHISKER FOR BOXPLOTS

$$Whisker_{lower} = e^{\ln(Q1)-1.5*\ln(\frac{Q3}{Q1})}$$

Where:

Q1 is the first quartile, also known as the 25th percentile;
Q3 is the third quartile, also known as the 75th percentile;
ln is the natural logarithm; and
e is the inverse log.

These components are illustrated in the example provided on Figure L-1. Individual results are overlaid on the boxplot as points to provide additional information about data variability.

## 2.2 SCATTERPLOTS AND STATISTICAL ANNOTATIONS

Scatterplots that show independent continuous variables along the x-axis and dependent continuous variables along the y-axis are used to explore the relationships among the Aquatic BERA data sets—such as the relationship between porewater metals concentrations and biological responses from sediment bioassays. Exploratory smoothing functions, such as locally weighted polynomial regression (LOESS), natural splines (NS), and linear model (LM) fits are presented on

scatterplots to provide visual guidance of data trends and underlying fluctuations. LM fits provide the overall linear trend between two variables, whereas LOESS curves fit weighted linear regression models to a window of data around each x value. The local polynomial fits give more weight to points near the x value and less weight on points far away from the x value. The final LOESS curve is obtained by connecting the locally fitted values. An NS is a piecewise polynomial function created from a set of functions to produce a mathematical model for a nonlinear curve. Unlike other types of spline models, NSs constrain the spline to be linear at the edges (beyond the boundary knots). Thus, they provide reasonable and generalizable fits across the range of x values.

LOESS and NSs can be helpful for exploratory purposes as these smoothing functions follow the underlying data fluctuations closely and are not constrained to have a straight-line fit. LOESS and LM fits are created in R using the argument `method = "loess"` and `method = "lm"` in the `geom_smooth()` function (see example on Figure L-2).

## 2.3 MOSAICS

Mosaic plots (also known as tile plots) are used throughout the benthic invertebrate community assessment sections of the Aquatic BERA to provide a visual overview of results from many analyses. Mosaic plots show an arrangement of colored rectangles, each representing a paired combination of variables. Results from analyses are represented by different colors, color intensity, and symbol overlays.

Two examples of mosaic plots include correlation matrices and percentile scores (Figure L-3). For the first example, pairwise correlations are computed for every combination of variables (e.g., all benthic invertebrate community metrics versus all bulk sediment metals analytes). Results are presented in mosaic plots where each rectangle is a variable-variable pair and the color and color intensity represent the strength and direction of the correlation. Statistical significance is represented by a point overlay or printed inside the rectangle (Figure L-3). In the second example shown on Figure L-3, mosaic plots are used to illustrate percentiles of the reference data distribution of benthic invertebrate community metrics for each sample location. For these graphics, each row represents a different sampling location and each column indicates a different community metric. The rectangles are filled with a color scale that indicates the percentile or percentile category of the community metric at that location.

It is important to remember that, as with any data visualization tool that divides continuous random variables into discrete bins, patterns may be sensitive to binning decisions. Most of the mosaic plots in the Aquatic BERA use color scale to represent the continuous variable (e.g., correlation coefficients). The only mosaic plots included in the Aquatic BERA that rely on decisions regarding discrete bins illustrate the reference envelope categories consistent with the Phase 3 Sediment Study Quality Assurance Project Plan (TAI 2019) and U.S. Environmental Protection Agency's (EPA) September 2020 Benthic Macroinvertebrate Discussion Summary Memo (USEPA 2020). For both bioassays and community metrics, the reference envelope percentile breakpoints are equivalent to the 5th percentile, 20th percentile, and 50th percentile.

# 3. EXPLORATORY AND CONTEXTUAL ANALYSES

This section describes the common statistical tools that were used to explore the Aquatic BERA data sets. This includes the ordination techniques (e.g., PCA) that were employed both for data exploration and to create composite variables for more advanced modeling and contextual analyses for further exploring factors related to benthic invertebrate community structure in Operable Unit (OU) 1 and upstream reference areas.

## 3.1 PRINCIPAL COMPONENTS ANALYSIS

PCA is used throughout the benthic invertebrate community assessment in the Aquatic BERA to understand patterns in complex multivariate data. This section provides an overview of the PCA method, a description of the outputs used in the Aquatic BERA, and a brief overview of the implementation.

PCA is a data reduction method that aids understanding of complex patterns in multivariate data. The aim of the analysis is to express covariation in many variables in a smaller number of composite variables. The PCA accomplishes this by seeking the strongest linear correlation structure among variables. The basic result is the variance explained by each axis (eigenvalues) and linear equations (eigenvectors) that combined the original variables. The principal component (PC) eigenvectors are the relative weight of each variable for calculating the PC scores and the PC loadings represent the correlation of the individual variable with the PC scores. Each object is projected into an ordination space using the eigenvector and the data matrix. For additional discussion of the PCA analysis including the specifics of the calculations, see McCune and Grace (2002).

A carefully constructed PCA captures most of the information in the data and distinguishes unique fingerprints. Often the first few PCs capture most of the information in the data set.

### 3.1.1 PCA METHODOLOGY

Executing PCA analyses includes making multiple decisions about the data handling and calculations. This section documents those decisions and provides a brief rationale for the approach.

As with all analyses in the Aquatic BERA, all nondetected chemistry results were substituted with the reported detection limit or, in the case of results with a "U*" validation qualifier, the detection limit value reported by the analytical laboratory. For the purposes of the PCA, analytical parameters having less than 80 percent detected values were removed from the PCA. Due to the uncertainty associated with use of replacement values to represent nondetected results, retaining the replacement values adds noise to the data set. Removal of parameters with a relatively high percentage of nondetects tends to reveal clearer results because it eliminates the noise created by using replacement values.

The natural log or logit transformation was applied where appropriate to strengthen the linear relationships between variables and reduce the influence of outlying values. Correlation matrices and scatterplots using all pairwise, complete data were examined to inform transformations for all

PCAs but were not necessarily reported for all analyses unless they were relevant to the risk questions in the Aquatic BERA.

Next, all variables were standardized to a mean of zero and variance of one. This is done by replacing each measured variable value with a new value that indicates the number of standard deviations (SDs) between the measured value and mean of the initial variable. PCA uses the eigenvalue decomposition of the variance-covariance matrix. However, when variability differs greatly across the initial variables, it is important to standardize the data. Standardization provides equal weighting of all variables in the analysis and prevents variables with large variances from dominating the results.

Once the data standardization is complete, the variance-covariance matrix is calculated on the standardized variables. Using the standardized data is equivalent to running the eigenvalue decomposition on the Pearson correlation matrix.[2] Pearson correlation coefficients summarize linear relationships and are sensitive to outliers.

Finally, matrix algebra was used to find the eigenvectors and eigenvalues of the correlation matrix (i.e., the eigenvalue decomposition was carried out on the standardized data). The fraction of the total information about the data set contained in each PC is found by dividing the eigenvalue for that PC by the sum of the eigenvalues for all PCs. It is common practice to discard any PC with an eigenvalue less than 1 (Kaiser 1960), as they explain less information than a single initial variable. This practice eliminates unimportant PCs that are driven by noise or multivariate outliers and allows a focus on the main patterns in the data. The eigenvalues were examined using scree plots (Section 3.1.2).

To visualize the PCA results, the PC scores were calculated for each sample. They indicate where a sample falls along the new axis in the projected space. For example, a sample with a high PC score is concurrently high in initial variables that load positively on that PC and low in the initial variables that load negatively on that PC. In general, metals that have loadings above 0.4 on a PC are considered important enough to be used as part of the interpretation of that PC as they have at least a moderately strong correlation with the component (Guilford 1956).

A customized R function was written to run all PCAs. This function uses the same calculations as the base R function stats::prcomp and produces the same results. The only difference is that the extraction of PCA results (loadings and scores) are simplified for downstream visualizations and outputs.

### 3.1.2 PCA VISUALIZATIONS AND OUTPUTS

The PC scores were assessed and interpreted using scree plots, PC loading plots, and PC biplots, all of which are included in the outputs for all PCAs in the Aquatic BERA analyses (Figures L-4a–c). These graphics help to identify similarities and differences in sample characteristics. The PC scores were also used in subsequent analyses, such as investigating relationships with benthic

[2] Positive correlation means that when x increases, y tends to also increase. Negative correlation means that when x increases, y tends to decrease. A correlation of 0 means that x and y fluctuate independently of one another. If the original data set contained n variables, then the correlation between variables can be displayed in an n by n matrix. The terms along the main diagonal are 1, the terms on the off diagonal are the correlations.

invertebrate community metrics and composite variables representative of habitat and metals exposure parameters.

**Scree plots** present the proportion of variance explained by each PC (i.e., its normalized eigenvalue) as well as the cumulative proportion of variance explained as more PCs are included (i.e., the sum of the first *k*-normalized eigenvalues sorted from largest to smallest). As *k* increases, each PC captures successively smaller portions of the overall variance of the data so that if all *k* PCs are included in the final solution, 100 percent of the overall variance is captured (Figure L-4a). A meaningful interpretation of the data can often be conveyed by including only the first few PCs. The point of inflection or "elbow" of the curve in the scree plot can be helpful for indicating the number of PCs to retain for further analysis (Cattell 1966), especially if there is an obvious point of diminishing returns in terms of variance explained by including additional PCs.

**PC loadings plots** (Figure L-4b) show the composition of PC scores and aid in their interpretation. Each loading is a transformation of the eigenvectors and can be interpreted as the correlation between the initial variable and the PC. The loadings plot is a bar chart displaying the loadings sorted from minimum to maximum that can be helpful when interpreting the PC scores.

**PC biplots** are two-dimensional scatterplots of the PC scores with PC loadings indicated as vectors (e.g., Figure L-4c). They are useful for visualizing and interpreting the predominant patterns captured by the key PCs. They can reveal clusters of similar samples and identify outliers.

Each point on the biplot represents a sample and can be color-coded according to important aspects of the data to facilitate further understanding. For instance, in Figure L-4c, color is used to indicate an area of the Site from which the sample was collected.

The vectors on the biplot indicate the strength and direction of the correlation between the original variables and the PCs being plotted. Longer vectors represent relatively stronger loadings.

- Initial variables whose vectors point to the right and up are positively correlated with both PCs.
- Initial variables whose vectors point to the right and down are positively correlated with the PC on the x-axis, and negatively correlated with PC on the y-axis.
- Initial variables whose vectors point to the left and down are negatively correlated with both PCs.
- Initial variables whose vectors point to the left and up are negatively correlated with the PC on the x-axis and positively correlated with the PC on the y-axis.

In the chemical context, samples that plot closely together in the biplot have similar chemical fingerprints. When looking for data clusters, it can be useful to explore three-dimensional plots of the PC scores, but these are most useful when used interactively where the data cloud can be spun around to reveal three-dimensional clusters. Typically, a set of two-dimensional biplots are used. For instance, three key PCs would result in three biplots: PC1 with PC2, PC1 with PC3, and PC2 with PC3.

### 3.1.3 PCA IMPLEMENTATION

The groups of parameters that were run for each PCA varied depending on the analysis question being addressed in the Aquatic BERA. The outputs were used both descriptively as exploratory tools (e.g., to identify clusters of samples with similar PC scores) and to generate key PCs that could be used in subsequent predictive modeling (e.g., Spearman rank correlations with benthic invertebrate community metrics in LOE 5).

## 3.2 NONMETRIC MULTIDIMENSIONAL SCALING

McCune and Grace (2002) declare NMDS as "the most generally effective ordination method for ecological community data and should be the method of choice." Minchin (1987) agrees, referring to NMDS as the most robust unconstrained ordination method in community ecology. Similar to PCA, NMDS is an ordination technique that aims to represent multidimensional data in a smaller dimension so that it can be easily visualized to reveal underlying structure and patterns in the data. Rather than ordination based on the correlation of all pairs of samples, as done with PCA, NMDS works with pairwise dissimilarities. In ecology, Bray-Curtis dissimilarity is often used. It is the ratio of the number of species in common to total number of species and is an appropriate measure of the dissimilarity of benthic invertebrate community composition.

In the ecological context, the original data indicate the number of taxa in each sample, or community. Typically, there are many taxa of interest. Thus, with one axis for each taxon, the data fall in high-dimensional space. NMDS attempts to reduce the dimensionality while retaining the original distance between samples, or communities. By using a dissimilarity measure such as Bray-Curtis, rather than Euclidean distance, the problem of samples with similar numbers of taxa being treated as similar even though the actual identity of the taxa differ is avoided.

NMDS is well suited to data that are non-normal or are on arbitrary or discontinuous scales such as taxonomic count data. Clarke (1993) provides an excellent summary of the advantages and uses of NMDS, especially in the context of assessing the composition of ecological community data.

PCA was an appropriate ordination technique for environmental chemistry data because the frequency of nondetected results was relatively low, data can easily be transformed to approach multivariate normality, and interpretation of PCA is relatively straightforward. In contrast, benthic invertebrate community data comprises species (or family) counts and the presence of rare taxa often makes it difficult to meet the assumptions of PCA even when the data are transformed.[3]

NMDS works with ranks to find a set of coordinates in a small dimensional space that minimize stress. Stress represents the difference between distance in the reduced space and distance in the full dimensional space. This helps avoid the assumption of linear relationships among variables and is especially helpful when community data has some rare species that are not well represented across samples.

[3] The presence of rare species in a community results in columns with many zeros, which can violate the assumptions of multivariate normality in PCA.

NMDS was implemented using the `vegan::metaMDS()` function in R (Oksanen et al. 2020). This function automates many of the recommended steps of NMDS including transformation of the data, multiple random starts to avoid getting trapped in local optima, scaling of the final results and calculation of species scores. In general, solutions with the fewest numbers of dimensions that have low stress (an inverse measure of fit to the data)[4] are preferred. As the number of dimensions increase, the stress value decreases but interpretability of the dimensions becomes much more difficult. A high-level summary of the steps in the analysis is provided below.

1. A sample-by-family matrix was created with species counts for each sample location (rows) and taxonomic family (columns). Higher taxonomic groupings were used when organisms could not be identified to the family level.
2. The sample-by-family matrix was converted to a dissimilarity matrix using Bray-Curtis coefficients. Bray-Curtis dissimilarity is commonly used for ecological data. It measures dissimilarity as the ratio of the number of species in common for two samples compared to the total number of species for the two samples. It can handle a large proportion of zero counts and will not consider species absence as similar (McCune and Grace 2002).
3. Because the abundance values were large (abundance ranged from 1 to 1,400 for some taxonomic groups), standardization was applied to improve results. The recommended Wisconsin double standardization and square root transformation were applied to both the abundance and relative abundance data sets.
4. NMDS was run using the `vegan::monoMDS` engine, a two-dimensional solution with at least 20 random starts. A solution was reported when at least two convergent solutions were found for a given number of dimensions. If at least two convergent solutions were not found, the number of dimensions was increased by one. The result with the lowest dimensions and lowest stress was reported as the final solution.
5. The final NMDS results were auto-scaled to center the origin of the NMDS scores to the average of the axes and improve readability of the biplots. Family scores were calculated as weighted averages as described in the `metaMDS` documentation (Oksanen et al. 2020).

The NMDS scores were plotted on x-y scatterplots that show all rotations of the final solution (e.g., NMDS axis 1 versus NMDS axis 2, NMDS axis 2 versus NMDS axis 3, and NMDS axis 1 versus NMDS axis 3). These biplots included multiple visual overlays including 95 percent confidence ellipses that use the standard error. They visually demonstrate the similarity in benthic invertebrate community composition among groups of interest.

Vector overlays were also created. Habitat variables, measures of diversity, tolerance, and abundance were fit onto the ordination axes using `envfit()` in the `vegan` package. Variables that had a significant correlation[5] to the NMDS axes are shown. The length of vectors represents the strength of association. Examples of an NMDS visualization are provided on Figure L-5.

[4] Values less than 0.05 represent a good fit, less than 0.1 is a fair fit, and greater than 0.2 is a poor fit.
[5] Significance was shown using a permutation procedure and a significance cutoff of 0.05.

## 3.3 INDICATOR TAXA ANALYSIS

The contextual analyses of the benthic invertebrate community data included an indicator taxa analysis, also referred to as multilevel pattern analysis. Indicator taxa analysis aims to identify which taxa are more often in one group of samples versus another. What distinguishes an indicator taxa analysis from other statistical methods evaluating taxa abundance is that it detects and describes the value of different taxa for indicating the overall environmental conditions. For example, if the goal is to use a species matrix to determine what ecosystem the samples came from, the presence of a grass species would not necessarily indicate that a sample was taken from a grassland because grasses are ubiquitous across multiple ecosystems. In contrast, the presence of a redwood tree can indicate that a sample was taken from a redwood forest since this species is found almost exclusively in that ecosystem.

Indicator taxa analysis combines the information on the concentration of taxa abundance in a particular group and the faithfulness of occurrence of a taxon in a particular group. A perfect indicator of a particular group should be faithful to that group (i.e., always present like the redwood example above). It should also be exclusive to that group, never occurring in other groups. The Dufrane and Legendre's indicator analysis uses these standards to calculate indicator values (IVs) for each species in each group (De Caceres and Legendre 2009; De Caceres et al. 2010; Dufrane and Legendre 1997).

Indicator taxa analysis was used in the contextual analyses of the benthic invertebrate community data to determine if some taxa were associated with certain habitats or sampling areas. Indicator taxa (e.g., those adapted to sandy conditions) identified through this analysis might help inform us about environmental conditions, determine the habitat preferences of particular taxa, and predict diversity of other species. All sampling locations (including reference) were included in this analysis regardless of the target strata and/or measured sand in the sample.

Indicator taxa analysis is calculated as follows:

1. Community samples are categorized by area based on the three OU 1 areas of interest (AOIs) in the Site as well as all upstream reference areas split between locations from the free-flowing reach and lacustrine locations from Lower Arrow Lake. Community samples from each AOI were also grouped based on substrate parameters—samples where substrate is more or less than 50 percent medium, coarse, and very coarse sand; estimated slag is more or less than 10 percent; excess simultaneously extracted metals are more or less than the median value; or samples where porewater copper is below or above its benchmark. For each taxon observed, the proportional abundance was calculated in a group relative to the abundance of that taxon in all groups. These proportional abundances are displayed as a percent.
2. The proportional frequency of the taxa in each group (i.e., the proportion of sample units in each group that contain that taxon) is also calculated and expressed as a percent.
3. The two proportions calculated in steps 1 and 2 are combined by multiplying them to produce the IV for each taxon in each group. Because the terms are multiplied, both the proportional abundance and frequency must be high to have a high IV score.
4. The statistical significance of the highest IV is evaluated by randomly reassigning taxa units to groups 1,000 times and recalculating the IV each time. The p value (i.e., the probability of a

Type I error) is the proportion of times that the IV from the randomized data set equals or exceeds the IV from the actual data set. The null hypothesis is that the IV is no larger than would be expected by chance (i.e., that the species has no IV).

Taxa with $p < 0.05$ were considered strong indicators of a specific AOI group meaning they were significantly more abundant in a specific AOI relative to the others. Indicator taxa analysis was performed using the `multipatt()` function in the `indicspecies` v. 1.7.9 R package (De Caceres and Legendre 2009). The taxa abundance matrix and vector of location assignments were used. Indicator taxa analysis was conducted for each taxon independently and results were summarized for all taxa that were significant indicators of a particular sampling area.

# 4. STATISTICAL APPROACH FOR THE BIOASSAY ANALYSES

This section describes the statistical approaches used for LOE 4 (bioassay and chemistry) in the assessment of the benthic invertebrate community ecological assessment endpoints. To address LOE 4, bivariate concentration-response models were developed for each bioassay endpoint (e.g., survival, growth, specific growth rate, and reproduction) using sediment and porewater chemistry measurements including the following parameters as predictors in the bivariate modeling:

- Concentrations of metals in sediment and porewater;
- Percent slag;
- Probable effects quotients for metals in sediment;
- Simultaneously extracted metals minus acid volatile sulfides; and
- Outputs from metals bioavailability models (e.g., Copper Biotic Ligand Model).

The following sections describe the statistical methodology used to develop concentration-response models in the LOE estimation in Section 4.1.4 in Volume 2 and Section 4.1.3 in Volume 3 of the Aquatic BERA.

## 4.1 MODELS USED

Two bivariate model types were fit to evaluate the relationship between metals concentrations and toxicity: sigmoidal and simple linear models (Figures L-6 and L-7). Although both model forms were fit, sigmoidal models were prioritized given their biological relevance as discussed below.

### 4.1.1 SIGMOIDAL MODELS

Sigmoidal models are commonly used to characterize bioassays because biological responses to exposures to toxic substances usually exhibit a nonlinear relationship. Sigmoidal models produce S-shaped curves that are characterized by a lower and upper asymptote, and a curve connecting the two. The EPA Toxicity Relationship Analysis Program uses sigmoidal curves to describe concentration-response relationships from toxicity tests (Erickson 2015).

The sigmoidal model is defined as (Equation 3):

#### EQUATION 3: FOUR-PARAMETER SIGMOID MODEL

$$y = c + \frac{d - c}{1 + \exp\{-b[\ln(x) - \ln(e)]\}}$$

Where:

ln(x) is the natural log transformed metal concentration;
exp() is the inverse of the natural log;
d is the upper asymptote;
c is the lower asymptote;

b is the slope at the inflection point;[6] and
e is the value of x, which gives a response halfway between c and d.

For the bioassay concentration-response modeling analyses, a three-parameter sigmoidal model (LL3) is used since the lower asymptote can be set to zero to represent the case where no organisms survive, grow, or produce biomass or offspring. Thus, the LL3 is a special case of Equation 3 (Equation 4, Figure L-6):

EQUATION 4: LL3

$$y = \frac{d}{1 + \exp\{-b[\ln(x) - \ln(e)]\ \}}$$

The LL3 is symmetric about the inflection point and assumes the underlying biological relationship is sigmoidal. The upper asymptote is allowed to vary to accommodate normalized values that may exceed 100. These model settings prevent model predictions less than zero. Parameters estimated include the slope, upper asymptote, and inflection point that corresponds to the effects concentration (EC) at which a 50 percent reduction in the biological response relative to the upper asymptote (EC50) is observed.

All LL3s were fit using the `drm()` function in the `drc` v. 3.0-1 R package. ECs were computed using the `ED()` function in the `drc` v. 3.0-1 R package (Ritz et al. 2015). EC20s[7] (and their corresponding y-intercepts, and confidence intervals) were calculated for all models as a conservative estimate of concentrations that evoke a biological response. No data transformations were applied to the bioassay endpoints while the metals concentrations were transformed as described in Appendix C (Attachments C1 through C3).

To fit LL3s, the underlying data must support a sigmoidal curve relationship. Data limitations may affect the accuracy of model fits, prevent model convergence, and decrease the ability to estimate model parameters accurately. Data limitations may include:

- Restricted range of chemistry responses so that the full range of biological responses are not observed;
- Insufficient number of points lying in the middle portion of the curve that are important for validating the slope estimates; and
- High leverage point(s).

LL3s require careful visual examination as the concentration-response relationships may be driven by high leverage points and actual model fits may not be meaningful. LL3 validity and acceptability were examined during the model selection phase (see Section 4.2 of this appendix for details). This includes a review of the directionality of the concentration-response relationship and a check to determine whether estimated EC20s were within the ranges of observed concentrations and whether all LL3 parameters were estimable.

[6] In these models, the inflection point is functionally equivalent to the EC50 or concentration at which a 50 percent reduction in the biological response relative to the upper asymptote is observed.
[7] In this application, EC20s are concentrations at which a 20 percent reduction in the biological response relative to the upper asymptote is observed.

For completeness, the following parameters are reported for all LL3s regardless of whether they were the final models selected for the particular analyte-endpoint pair:

- The range of observed x values;
- The slope (b), upper asymptote (d), and inflection point (e);
- The endpoint response that corresponds to the EC20 and the concentration (and 95 percent confidence interval) that corresponds to the EC20 estimate; and
- The quasi coefficient of determination ($R^2$) (defined as the squared correlation between observed and model predicted values).

### 4.1.2 SIMPLE LINEAR MODELS

Simple linear relationships were considered a viable model for cases where the data did not support a sigmoidal relationship. The LM assumes a linear relationship between bioassay endpoint and chemistry concentrations, meaning a constant change in chemistry leads to a constant change in the bioassay endpoint.

The model form is defined as follows (Equation 5, Figure L-7):

#### EQUATION 5: LINEAR REGRESSION MODEL

$$y = \alpha + \beta x$$

Where:

$\alpha$ is the y intercept;
$\beta$ is the slope; and
$x$ is the natural log transformed chemistry concentration.

The LM assumes a straight-line relationship between endpoint and chemistry throughout the range of chemistry concentrations. Due to the restriction in range, the shape of the model may not be fully apparent. For instance, there may be a straight-line relationship for the given range of x values but if lower concentrations were observed, the upper asymptote may have been quantifiable. Similarly, the inflection point and lower asymptote of a sigmoid model would only be quantifiable if higher concentrations were observed in the study. While linear concentration-response models have limited generalizability outside of the study, USEPA (2000) does consider it a valid model form for concentration-response modeling. Since no constraints are imposed on the model parameters, model predictions may result in biological responses that are greater than 100 percent or less than 0 percent (i.e., predicted responses that do not make biological sense). Model predictions and interpretation is limited to within the observed chemistry concentrations. In addition, LMs are sensitive to outliers and high leverage points and these influential points may modify the true underlying relationship, (i.e., y intercept and slope estimates).

For completeness, the following parameters are reported for all LMs regardless of whether they were the final models selected for the particular analyte-endpoint pair:

- The range of observed x values;
- The slope, its standard error and p value, and direction (positive or negative);

- The intercept and its standard error; and
- The quasi $R^2$, which in the case of LMs is equivalent to a traditional $R^2$.

### 4.1.3 MEASURING GOODNESS OF FIT

As a goodness-of-fit metric, a quasi $R^2$ was used to compare the goodness of fit across the nonlinear and LMs. $R^2$ is a common measure of goodness of fit in linear regression and is defined as the proportion of variance explained by the regression fit. In other words, it is a measure of the level of agreement between an observed set of values and a set wholly or partly derived from a model of the data.

Due to different underlying assumptions for linear and nonlinear regression, $R^2$ is not a valid goodness-of-fit measure as has been demonstrated by simulation studies in Spiess and Neumeyer (2010). As an alternative, a quasi $R^2$, defined as the squared correlation between model predicted and observed values, was used as a goodness-of-fit measure. This quasi $R^2$ measure is restricted to be between 0 and 1, and in the scenario of a perfect fit would exhibit a quasi $R^2$ value of 1. The quasi $R^2$ was used to provide a familiar goodness-of-fit metric for the LL3s. The quasi $R^2$ from LL3s and $R^2$ from the LMs are used to inform the scoring of model strength in the weight of evidence step. Additional metrics were used to determine model validity, strength, and reliability (e.g., direction of the slope and estimates of model accuracy).

## 4.2 DECISION FRAMEWORK FOR MODEL SELECTION AND VALIDATION

An objective decision framework was developed to validate and select the final bivariate model for each predictor-endpoint data set that best characterized the concentration-response relationships (see Figure 4-1 from Volume 1 of the Aquatic BERA for an overview). All model types (LL3 and LM) were fitted to the data and a single "best" model per analyte-response pair was identified in the LOE estimation (Section 4.1.4 in Volume 2 and Section 4.1.3 in Volume 3 of the Aquatic BERA). The decision framework consisted of five distinct steps:

1. Model development
2. Model validation
3. Screening for biological relevance
4. Flagging models
5. Model selection

### 4.2.1 STEP 1: MODEL DEVELOPMENT

To be as inclusive as possible in the concentration-response modeling, all data sets with at least 50 percent detected values in the predictor variable were advanced to the model development stage. This differed from the detection frequency cutoff used for the PCAs (Section 3.1.1) because the uncertainty associated with reduced detection frequency can be addressed individually for each parameter where it is applicable. Where predictor variables were reported as detected values in at least 50 percent of samples in the bioassay data set, models were initially screened using Spearman rank correlation consistent with direction from USEPA (2022, pers. comm.). Bioassay

response endpoints and predictor variable pairs that met both the following criteria were advanced to the subsequent model selection steps:

- P value < 0.05
- |r| > 0.3

An attempt was made to fit both model types (LL3 and LM) to the response/predictor variable pairs that were retained following the Spearman rank correlation screen as described in the subsequent steps.

### 4.2.2 STEP 2: MODEL VALIDATION

All models were screened to make sure they met minimum criteria for statistical validity. For the LL3s, commonly used statistics for evaluating models (e.g., p values and $R^2$ values) are not reliable as the statistical significance of LL3 parameter estimates do not provide direct indication of a significant or valid LL3 relationship.

Instead, all of the following criteria must have been met to determine if the LL3 was statistically valid:

- The model converged on a solution.
- All three model parameters (slope, inflection point, and upper asymptote) were estimated in the model fitting step.
- Predicted values (e.g., EC20) were estimated within the range of observed x values.

Careful inspection of the data sets and models that failed these screening criteria (examples provided on Figure L-8) were poorly suited for sigmoidal models for any number of reasons including, but not limited to:

- Visual plots of the data set with a fitted LOESS line indicated that there was no underlying sigmoidal shape to the data;
- Restriction on the x-range prevented the fitting of a sigmoid model;
- High variability in the endpoint response prevented reliable model development; and
- Fitted models produced nonsensical results with confidence bands or predicted values well outside of reasonable ranges.

LMs were screened using more standard criteria. Valid LMs had slopes significantly different from zero ($p$ less than 0.05) and $R^2$ greater than 0.09 ($|r| \geq 0.3$). These numeric cutoffs are consistent with common statistical convention and are conservative in that they include (rather than exclude) weak models.

### 4.2.3 STEP 3: SCREENING FOR BIOLOGICAL RELEVANCE

Theoretically, higher chemical concentrations result in poorer biological performance. Thus, slopes for all LL3 and LMs were screened for the directionality of the slopes (b term for the LL3 and β for LMs). Negative slopes were consistent with theoretical concentration-response models, whereas models with positive slopes do not indicate a toxicological response. Only models whose fitted slopes were negative were advanced in the model selection process.

### 4.2.4 STEP 4: FLAG MODELS

While a model may be statistically valid and biologically relevant, it can also have limited reliability or relevance. Some preliminary measures of reliability and relevance were flagged to interpret selected models. For example, LL3s with high leverage points were flagged for further evaluation because they may produce predicted values that are not biologically meaningful. Initial review of the fitted LL3s showed that a single data point often determined the LL3 fit and created a spurious LL3 relationship. To flag such cases, Cook's distance (D) was calculated for all sample points and used as a measure to identify highly influential points. Rough, rule-of-thumb metrics exist to determine the cutoff for large D such as D greater than 4/n, D greater than 1, or D values much larger than the rest of the D distribution (Cook 1977). Histograms of D values were examined for all endpoint and analyte combinations and D greater than 3 was used to flag samples as having high leverage on the model fit. Smaller cutoffs were also examined; however, many samples with D less than 3 were flagged but did not appear to drive the spurious LL3 relationships. A cutoff of 3 captured most cases where LL3 relationships appeared spurious due to a single sample.

### 4.2.5 STEP 5: MODEL SELECTION

All remaining statistically valid models with slopes in the biologically relevant direction were entered into the final model selection step. Two guiding principles were used to select models: that of highest biological relevance and that of statistical parsimony. The LL3 was considered the most biologically relevant model because the sigmoid shape is best aligned with the theoretical expectation of the biological response (Erickson 2015). In contrast, the LM could be preferred based on parsimony if it explained the data as well as the LL3. To resolve the tension between these two principles, an objective model comparison metric (Akaike Information Criterion [AIC]) was used to select from among models that were "tied" for the best model. If the LL3 was among the models tied for the best model, it was preferentially selected before the LM.

The AIC is a commonly used measure for selecting among biologically relevant models (Burnham and Anderson 2002). AIC is computed based on the log likelihood and is applicable across linear and nonlinear regression models. It therefore overcomes the limitations of other model comparison techniques that require models to be nested within each other for them to be compared. AIC measures the relative fit of each proposed model and provides an estimate of the information lost by using a model to represent the observed data. It provides a measure of the trade-off between the goodness of fit of the model and the simplicity of the model. AIC was used to objectively compare the LL3s and LMs to each other.

AIC is a relative measure used to compare among proposed models. Hence, if all models are a poor fit, it indicates which of these "poor" fitting models has the best fit to the data; it does not indicate the absolute quality of the fit. The model with the lowest AIC value is typically considered to have the highest probability of explaining the data. However, a general rule of thumb (Burnham and Anderson 2002) suggests that models associated with AIC values within two points provide relatively similar goodness of fit.

All models that were within two points of the minimum AIC value were considered equally valid for the best solution. Based on the principle of biological relevance, the LL3 was preferentially selected among these best models followed by the LMs.

## 4.3 SUMMARY OF FIGURES AND PRESENTATION OF FINDINGS

The decision framework was executed for all bioassay endpoint and parameter combinations. A summary of findings has been compiled into mosaic plots in the Aquatic BERA that indicate the final model form selected (indicated by color where blue = LL3 selected, orange = LM selected, and gray = no LL3 or LM selected as depicted in Figure L-3A). Black dots are used to indicate LL3s that had a high leverage value with D greater than 3. The color intensity scale reflects the goodness of fit using the quasi $R^2$.

LL3s and LMs for all endpoint-predictor combinations are reported in tables in the Aquatic BERA. Scatterplots with a line indicating the best model fit for each analyte (e.g., porewater metals) and bioassay response are provided in Figures L-8 and L-9.

## 4.4 MODEL DIAGNOSTICS USED IN WEIGHT OF EVIDENCE SCORING

The model selection process described in Section 4.2 is the basis for the LOE estimation for LOEs 4 (Volume 2) and 3 (Volume 3) of the Aquatic BERA. As described in Volume 1 (Section 4.3.3), the evidence within each LOE is assessed for strength, reliability, and relevance in the weight of evidence assessment for benthic invertebrate communities. For the concentration-response modeling with sediment bioassay data, models are scored based on their strength and reliability based on the following model parameters:

- The EC20 for the final selected model (per Step 5 above) is used as the exposure threshold, above which samples would be predicted to be "toxic."
- Bioassay response value that corresponds to the EC20 (see Figure L-9) is used as the response threshold below which responses are considered "toxic."

Based on these two modeled values, the following diagnostics are calculated for reliability scoring:

- **Accuracy rate**: Samples with exposure values below the EC20 and with responses above the corresponding effect level (i.e., are in the top left quadrant of the scatterplots, see Figure L-9) are classified as accurate predictions of no toxicity. Samples with exposure values exceeding the calculated EC20 and with response below the corresponding effect level (i.e., are in the bottom right quadrant of the scatterplots, see Figure L-9) are classified as accurate predictions of toxicity. Samples in the other two quadrants (i.e., top right, with exposure values exceeding the calculated EC20 but not exhibiting "toxic" response, and bottom left, with exposure values not exceeding the calculated EC20 but exhibiting a "toxic" response) are classified as inaccurate based on the final model. The accuracy rate, therefore, is calculated as the count of all accurate samples divided by the total count of samples.
- **False positive** rate is calculated as the count of all false positive samples (i.e., the exposure value exceeds the calculated EC20, but the effect level is not "toxic") divided by the count of all samples with exposure values exceeding the calculated EC20.
- **False negative** rate is calculated as the count of all false negative samples (i.e., the exposure value does not exceed the calculated EC20, but the effect level is "toxic") divided by the count of all samples with exposure values not exceeding the calculated EC20.

Models with correlation values ($R^2$) > 0.16 (see Volume 1, Section 4.3.3) are scored as "uncertain" if the accuracy rate is less than 60 percent or either the false positive or false negative rate exceeds 40 percent. Models with correlation values > 0.16 are scored as "possible" if the accuracy rate is greater than 60 percent and both the false positive and false negative rates are less than 40 percent.

# 5. STATISTICAL APPROACH FOR THE BENTHIC INVERTEBRATE COMMUNITY ANALYSES

This section describes the statistical approaches used for LOE 5 in the assessment of the benthic invertebrate community ecological assessment endpoints. LOE 5 includes the concentration-response modeling of benthic invertebrate community metrics.

In the LOE estimation for LOE 5 in Section 4.1.5 of Volume 2 of the Aquatic BERA, co-located samples of benthic invertebrate community data, sediment and porewater metals chemistry, and sediment physical characteristics are examined using bivariate relationships. In the cases where significant relationships are found, multivariate analyses are conducted to refine the relationships and account for the influence of multiple variables on the benthic invertebrate community metrics. The following sections describe the data handling and bivariate and multivariate modeling methods that were used to answer the LOE 5 analysis question for OU 1 in Section 4.1.5 of Volume 2 of the Aquatic BERA.[8]

## 5.1 DATA HANDLING

General data management and data treatment of the 2019 Phase 3 study data prior to use in analyses is discussed within the Final Phase 3 Data Summary Report (TAI 2021). Additional data handling of 2019 Phase 3 data prior to use in statistical analyses is discussed below.

### 5.1.1 DATA TRANSFORMATIONS

Because the concentration-response analyses for the benthic invertebrate community metrics used Spearman rank correlations, no data transformations were required for the community metrics or the various predictor variables. However, transformations were applied, where appropriate, in the scatterplots used to facilitate exploration of relationships.

### 5.1.2 LOW FREQUENCY OF DETECTED VALUES

The effects of sediment and porewater chemistry on the benthic invertebrate community may not be accurately estimated if a chemistry variable has a low frequency of detection. Including a high proportion of nondetected chemistry values and treating them as detected ignores any uncertainty in these observations and the analysis will tend to underestimate the SD of the variables. Therefore, a detection frequency cutoff of 75 percent was used to maximize the number of detected observations and analytes for the regression analyses.

While this cutoff is higher than that the cutoff used for the concentration-response modeling for bioassays (Section 4.2.2), the 75 percent cutoff was selected based on an assessment of the data and best professional judgment. All sediment analytes were close to 100 percent detected, and, while slightly more variable, most porewater analytes were detected in nearly 100 percent of the

[8] Benthic invertebrate community data were only collected as part of the 2019 Phase 3 sediment study and are, thus, only available for OU 1. LOEs based on benthic invertebrate community data are not included in the Aquatic BERA for OU 2.

samples as well[9]. The detection frequencies for the following porewater analytes were below 75 percent in the Phase 3 data set and were, therefore, not included in the concentration-response modeling with the benthic invertebrate community metrics: beryllium, selenium, silver, and thallium. Of these, beryllium, silver, and thallium were not identified as chemicals of potential concern for benthic invertebrates in porewater. Selenium in porewater is identified as a chemical of interest and is evaluated qualitatively in the Aquatic BERA (see Table 2-8a in Volume 1). Therefore, no further concentration-response analyses were conducted with these analytes.

## 5.2 BIVARIATE CORRELATION MODELS

Correlation between benthic invertebrate community metrics and chemistry and habitat variables was evaluated using Spearman rank correlation (Equation 6).

### EQUATION 6: SPEARMAN RANK CORRELATION

$$r_s = \frac{cov(R(X), R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}}$$

Where:

$r_s$ is the Spearman rank correlation coefficient;
$R(X)$ and $R(Y)$ are the raw scores of variables *X* and *Y* converted to ranks based on value;
$cov(R(X), R(Y))$ is the covariance of the rank variables; and
$\sigma_{R(X)}$ and $\sigma_{R(Y)}$ are the SDs of the rank variables.

Models were considered valid and flagged for further analysis if the following conditions were met:

- The slope of the bivariate relationship is in the direction of potential impairment.
- The chemistry effect is statistically significant (p less than 0.05).
- The correlation between benthic invertebrate community metric and chemistry is at least 0.3, which is equivalent to $R^2 \geq 0.09$.

No *a priori* directionality was assumed in the regression of benthic invertebrate community metrics with the habitat metrics. This was because the objective for including habitat metrics in the bivariate analyses was simply to determine where there was significant association between the benthic invertebrate community and habitat metrics.

Scatterplots of benthic invertebrate community metrics against each chemistry variable, along with the fitted lines and exploratory LOESS fits (Section 2.2), were used to visually assess the strength and validity of the bivariate relationships.

Habitat PCs from the physical characteristics and general chemistry habitat PCAs were included as additional environmental variables.

[9] For the purposes of this analysis, results where U* validation flags were applied were considered detected values. Samples flagged with the U* validation qualifier refer to results reported by the analytical laboratory that were within a factor of 5 of the detected concentration in the corresponding equipment blank. In the Phase 3 field porewater data set, metals such as copper and zinc were detected at low but consistent concentrations in all equipment blanks.

## 5.3 MULTIVARIATE MODELING

Teasing apart which factors are most likely to be contributing to differences in benthic invertebrate community metrics at the Site is confounded by the fact that several of the chemistry and habitat parameters evaluated above are correlated with each other. The statistical power to distinguish between the influence of exposure to metals and differences in habitat, unrelated to the presence of metals, in the Phase 3 benthic invertebrate community data set is uncertain. Therefore, the following approach was used to compare the relative effects of metals exposure and habitat factors on benthic invertebrate community metrics.

PCA on all sediment, porewater, surface water, water chemistry, and habitat variables (e.g., grain size, elevation, distance to thalweg) was conducted using the methods described in Section 3.1.1. This provided a dissection of the habitat and metals parameters that tended to be found most frequently together. There were two iterations of this PCA process, one with reported concentrations of metals and one with proportional concentrations of metals within each sample type (sediment, porewater, surface water).

The PCs from these PCAs that accounted for up to 90 percent of the variance in the combined chemistry and habitat data set were then included in bivariate comparisons with benthic invertebrate community metrics as described in Section 5.2.

## 5.4 PERCENT EPT MODELING

More than 50 percent of sample locations did not have pollution intolerant taxa (Ephemeroptera, Plecoptera, and Trichoptera [EPT]) present; therefore, percent EPT was converted to a Boolean variable based on whether those families were present (1) or absent (0) in the sample. The effects of chemistry on EPT were assessed using *t*-tests constructed to examine if there are differences in chemistry concentrations for samples where EPT is present compared to samples without EPT. Chemistry variables were flagged if the differences in concentrations were statistically significant (p less than 0.05) and the difference was in the expected direction (i.e., observing lower chemistry concentrations when EPT are present).

# 6. REFERENCE COMPARISONS

As part of the analyses for LOE 2 and LOE 3, bioassay response and benthic invertebrate community metrics from the Site were compared to those from upstream reference locations. These reference comparisons were conducted using two different approaches: (1) comparison to the distribution of values from the upstream reference locations, or reference envelope; and (2) comparisons of central tendency bioassay responses and benthic invertebrate community metrics from Site locations to the central tendencies from the reference locations.

In both types of reference comparisons with the Phase 3 data, an initial assessment was conducted to ensure that the reference samples are representative of the range of conditions in the corresponding OU 1 samples. This step was conducted to account for the potentially confounding effect of reduced bioassay performance in sediment samples from specific substrates (e.g., low total organic carbon [TOC]) that was observed in the Phase 3 bioassays (USEPA 2021). The following analyses were conducted to confirm which response endpoints were likely confounded by poor laboratory performance and, if so, which reference area samples should be removed from the reference area data set prior to the reference envelope and central tendency comparisons to reference:

1. For the reference and laboratory control samples only, the concentration-response modeling framework (Section 4.2 in this appendix) is applied to the Phase 3 bioassay response data with sediment TOC and pH as the exposure variables to determine whether bioassay response may be confounded by either parameter.
2. If no significant model is identified, bioassay response is not considered confounded by that exposure variable, and no adjustments are necessary to the reference area data set.
3. If a significant model is identified, upstream reference samples with sediment characteristics (e.g., TOC) less than the calculated EC20s in the models developed in Step 1 are identified for removal from the reference area data set for each bioassay response endpoint comparison. These are the samples where bioassay responses are meaningfully different from the remaining reference samples due to the confounding effects evident in the negative control samples.

## 6.1 REFERENCE ENVELOPES

The reference envelope analyses in the Aquatic BERA were conducted consistent with the Phase 3 Sediment Study Quality Assurance Project Plan (TAI 2019). As such, each Site sample is categorized based on bioassay results and benthic invertebrate community metrics relative to reference area results (e.g., less than 80th and less than 50th percentiles of bioassay reference sample results, and/or less than 50th and 5th percentiles of the benthic macroinvertebrate [BMI] reference sample results). Reference envelopes were developed using the percentiles of the reference distribution.

Site samples were compared to the percentiles of the reference sample distribution. The reference sample distribution was first calculated using the `ecdf()` function in R. The percentiles that were calculated from that distribution were dependent on the directionality of the metric or endpoint of interest (Table L-1). For all bioassay endpoints and benthic invertebrate community metrics where

lower values are expected to be indicative of potential impairment, the following percentiles were calculated for the reference data set: 5th, 20th, and 50th percentiles. For benthic invertebrate community metrics where higher values are expected to be indicative of potential impairment, the following percentiles were calculated for the reference data set: 95th, 80th, and 50th percentiles.

Each Site sample's value was equated to a percentile of the reference sample distribution. The Site samples' computed percentiles were then classified into bins based on each metric's direction of potential impairment. Data tallies were created for the number of samples below the 5th, 20th, and 50th percentiles or above the 95th, 80th, and 50th percentiles depending on the direction of potential impairment. Because this approach uses multiple percentiles of the reference distribution, it provides a more direct comparison of the Site and reference distributions.

Scatterplots of bioassay response or community metric value versus river mile were used to visually compare the biological response distributions in reference versus Site where horizontal dashed lines indicate the reference sample distribution's percentiles of interest. Mosaic plots were used to show the percentile bins, actual percentiles, and exceedances of each Site sample relative to the reference distribution.

## 6.2 CENTRAL TENDENCY COMPARISONS

The central tendency comparisons between Site and upstream reference locations were conducted consistent with guidance received from EPA in September 2020, in which EPA recommended sequential analyses starting with traditional hypothesis testing (USEPA 2020, pers. comm.). If no differences were determined, EPA recommended a subsequent test to assess whether the Site results were equivalent to reference with a pre-specified magnitude. Specifically, EPA's September 8, 2020, memo states:

> *Where the traditional approach does not reject the null hypothesis, the level of difference will be evaluated with a reverse hypothesis test, which will determine the level of equivalence. We anticipate that demonstration of relative differences up to 50% from the reference mean would be considered evidence of equivalence.*

This two-step process can be summarized in a schematic (Figure L-10). Note that the color coding associated with each of the outcomes of the sequential analyses are carried into the results tables in the main body of the Aquatic BERA. First, each bioassay response endpoint and benthic invertebrate community metric was assessed using traditional hypothesis testing to assess whether a significant difference in means between groups could be established. Where a significant difference is found with a *t*-test, then Site and reference are declared to be different.

Where a *t*-test did not indicate a significant difference between Site and reference sample means, the data were assessed using equivalence testing. Equivalency testing requires that the difference between the two groups is specified *a priori.*[10] Equivalency testing provides a p value that, when significant, indicates that the two populations are statistically equivalent within a specified margin of error (α was set to 0.05 for both the bioassay response and benthic invertebrate community metric data). Data sets without a significant equivalence test were considered uncertain and the

[10] This difference is referred to as the "effect size."

Site and reference sample populations were neither determined to be different nor equivalent. In all cases where differences were found between the Site and reference sample population, interpretation focused on the magnitude of that difference and its biological relevance.

Among the benthic invertebrate community metrics, total abundance, density, and counted blotted mass were log transformed to satisfy the assumptions of normality for all *t*-tests. No transformations were applied to the other benthic invertebrate community metrics or any of the bioassay endpoints.

### 6.2.1 TEST FOR SIGNIFICANT DIFFERENCE

The bioassay responses and benthic invertebrate community metrics were compared using a one-sided *t*-test to evaluate the following hypotheses on Figure L-11. The directionality of the one-sided test depended on the metric being evaluated.

No further evaluation was pursued for metrics that showed that Site was significantly different from reference (accept $H_A$ as determined by p less than 0.05). Metrics that did not show a significant difference in Site samples versus reference samples were advanced to the equivalency testing (accept $H_0$ as determined by p greater than 0.05).

### 6.2.2 TEST FOR EQUIVALENCY

Equivalency tests examine whether differences between groups were within bounds that are deemed as practically equivalent. If mean differences were beyond these bounds, then the two samples were deemed as not equivalent, or outside the acceptable equivalence range. In other words, the null hypothesis assumed that the two groups were different by at least δ (effect of interest). The one-sided null and alternative hypotheses for equivalency testing depend on the directionality of the metric and are provided on Figure L-11.

The equivalency bound[11] (δ) is set so that when differences between the two populations are not at least as large as δ, they are deemed too small to be of meaning (i.e., they are functionally equivalent). Equivalence bounds may be in the form of raw differences or standardized differences (Cohen's *d*). Cohen's d is the equivalence metric typically used to define a "meaningful" difference between two means. Cohen's *d* is defined as shown below (Equation 7).

#### EQUATION 7: CALCULATION FOR COHEN'S D

$$d = \frac{\mu_{Ref} - \mu_{Site}}{\sigma}$$

Where:

The numerator is the difference between the means between Site and reference; and

The denominator $\sigma$ is the pooled SD between the two groups.

A value of d = 0.2 is considered a small effect size, 0.5 is a medium effect size, and 0.8 is considered a large effect size. Statistical methods for testing equivalency followed Walker and Nowacki (2011).

[11] This is also described as the effect size.

The equivalence bounds were set at a standardized difference, Cohen's *d* of 0.5, which would indicate that the means of two groups differ by 0.5 (pooled) SDs, which is considered a "medium" effect size. The equivalence bounds were expressed in the original units of the metal concentration, bioassay endpoint, or BMI community metric by multiplying Cohen's *d* by the observed pooled SDs.

Percent EPT, due to the high number of zeros in the data, was tested for equivalency using a chi-squared test. In this test, the proportion of samples with percent EPT greater than zero was compared between reference and Site locations.

# 7. SOFTWARE

All statistical analyses and visualizations were performed using R Statistical Computing language (R Core Team 2019) unless otherwise specified. Packages that were used for specific statistical calculations that are not included as part of the R base package are specified and cited as appropriate. Most figures were produced using `ggplot2`, `tidyverse`, and associated packages.

# 8. REFERENCES

Burnham, K.P., and D.R. Anderson. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2nd ed. New York: Springer-Verlag.

Cattell, R.B. 1966. "The Scree Test for the Number of Factors." *Multivariate Behavioral Research* 1(2): 245–276.

Clarke, K.R. 1993. "Non-parametric multivariate analyses of changes in community structure." *Australian Journal of Ecology* 18: 117–143.

Cook, R. Dennis. 1977. "Detection of Influential Observations in Linear Regression." *Technometrics* 19(1): 15–18.

De Caceres, M., and P. Legendre. 2009. "Associations between species and groups of sites: indices and statistical inference." *Ecology* 90(12): 3566–3574. https://doi.org/10.1890/08-1823.1.

De Caceres, M., P. Legendre, and M. Moretti. 2010. "Improving indicator species analysis by combining groups of sites." *Oikos* 119(10): 1674–1684.

Dufrane, M., and P. Legendre. 1997. "Species assemblages and indicator species: The need for a flexible asymetrical approach." *Ecological Monographs* 67: 345–366.

Erickson, R. 2015. *Toxicity Relationship Analysis Program (TRAP)*. Edited by U.S. Environmental Protection Agency. Washington, DC.

Guilford, J.P. 1956. "The structure of intellect." *Psychological Bulletin* 53(4): 267–293. https://doi.org/https://doi.org/10.1037/h0040755.

Kaiser, H. 1960. "The Application of Electronic Computers to Factor Analysis." *Education and Psychological Measurement* 20(1): 141–151.

McCune, B., and J.B. Grace. 2002. *Analysis of ecological communities*. MjM Software Design, Gleneden Beach, OR.

Minchin, P.R. 1987. "An evaluation of relative robustness of techniques for ecological ordinations." *Vegetation* 69: 89–107.

Oksanen, J., F. Guillaume Blanchet, M. Friendly, R. Kindt, P. Legendre, D. McGlinn, P.R. Minchin, R.B. O'Hara, G.L. Simpson, P. Solymos, M. Henry, H. Stevens, E. Szoecs, and H. Wagner. 2020. *Vegan: Community Ecology Package R package version 2.5-7*.

R Core Team. 2019. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ritz, C., F. Baty, J.C. Streibig, and D. Gerhard. 2015. "Dose-Response Analysis Using R." *PLOS ONE 10* 12. https://doi.org/https://doi.org/10.1371/journal.pone.0146021.

Spiess, A.N., and N. Neumeyer. 2010. "An evaluation of R2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach." *BMC Pharmacology* 10. https://doi.org/https://doi.org/10.1186/1471-2210-10-6.

TAI (Teck American Incorporated). 2019. *Final Quality Assurance Project Plan for the 2019 Phase 3 Sediment Study*. Prepared for Teck American Inc. by ERM in association and consultation with Windward Environmental LLC, HDR, Exponent, and Parametrix, Inc.

TAI. 2021. *Final 2019 Phase 3 Sediment Study Data Summary Report.* Prepared for Teck American Inc., Spokane, WA, by ERM, Carpinteria, CA, in association and consultation with Windward Environmental LLC, HDR, Exponent, and Parametrix, Inc.

USEPA (U.S. Environmental Protection Agency). 2000. *Methods for measuring the toxicity and bioaccumulation of sediment-associated contaminants with freshwater invertebrates*. EPA/600/R-99/064. U.S. Environmental Protection Agency, Office of Research and Development.

USEPA. 2006. *Settlement Agreement for Implementation of Remedial Investigation and Feasibility Study at the Upper Columbia River Site*. June 2, 2006.

USEPA. 2020. *BMI Discussion Summary*. Memorandum from K. Cerise, U.S. Environmental Protection Agency, to K. McCaig, Teck American Incorporated, dated September 8, 2020.

USEPA. 2021. *Phase 3 Sediment Study 42-day Hyalella azteca Bioassay Quartz Sand Negative Control Performance Evaluation*. Prepared by CH2M HILL on behalf of U.S. Environmental Protection Agency.

USEPA. 2022. Personal communication (email to D. Mills, Teck American Incorporated, from B. Arthur, U.S. Environmental Protection Agency, dated March 4, 2022, regarding *EPA Draft Comments on the Phase 3 Tech Memo*). U.S. Environmental Protection Agency, Seattle, WA.

Walker, E., and A. Nowacki. 2011. "Understanding equivalence and noninferiority testing." *Journal of general internal medicine* 26(2): 192–196. https://doi.org/https://doi.org/10.1007/s11606-010-1513-8.

Wickham, H., M. Averick, J. Bryan, W. Chang, L. D'Agostino McGowan, R. Francois, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. Lin Pedersen, E. Miller, S.M. Bache, K. Muller, J. Ooms, D. Robinson, D.P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani. 2019. "Welcome to the Tidyverse." *Journal of Open Source Software* 4(43): 1686. https://joss.theoj.org/papers/10.21105/joss.01686#.

# TABLES

Table L-1
Directionality and Percentile Bins of Benthic Invertebrate Community Metrics
Upper Columbia River
Final Aquatic Baseline Ecological Risk Assessment
Appendix L

| Community Metric | Direction of Impairment | Bins |
|---|---|---|
| Total Abundance, Density, Counted Blotted Mass | Low = Impaired | ≤ 5th percentile; 5–20th percentile; 20–50th percentile; > 50th percentile |
| Shannon-Weaver Diversity, Species Richness | | |
| Predator Richness | | |
| % Dominant Taxon, Metals Tolerance Index | High = Impaired | ≥ 95th percentile; 80–95th percentile; 50–80th percentile; < 50th percentile |
| Hilsenhoff Biotic Index, % Chironomidae | | |

FIGURES

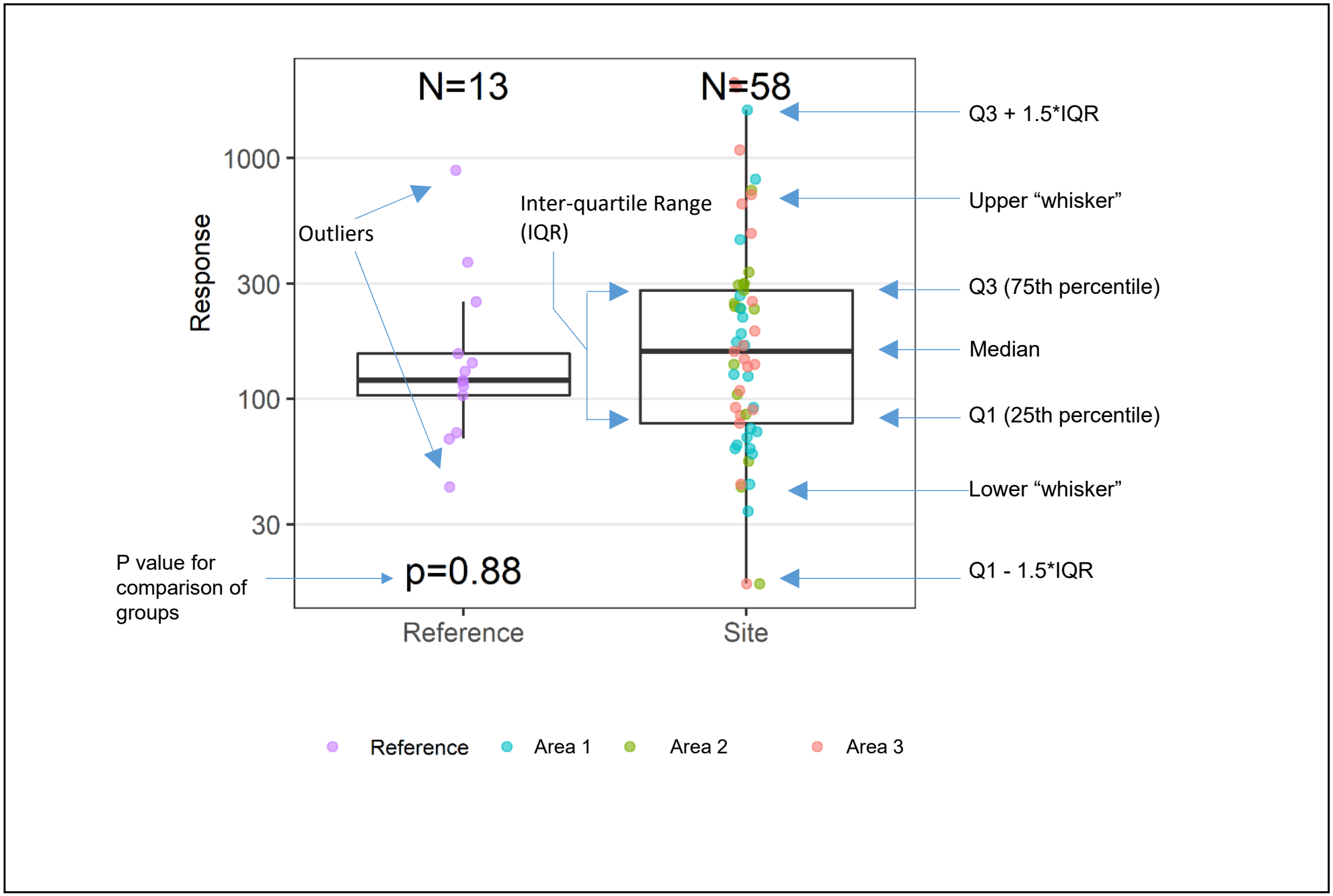



Figure L-1. Anatomy of a Boxplot

ERM

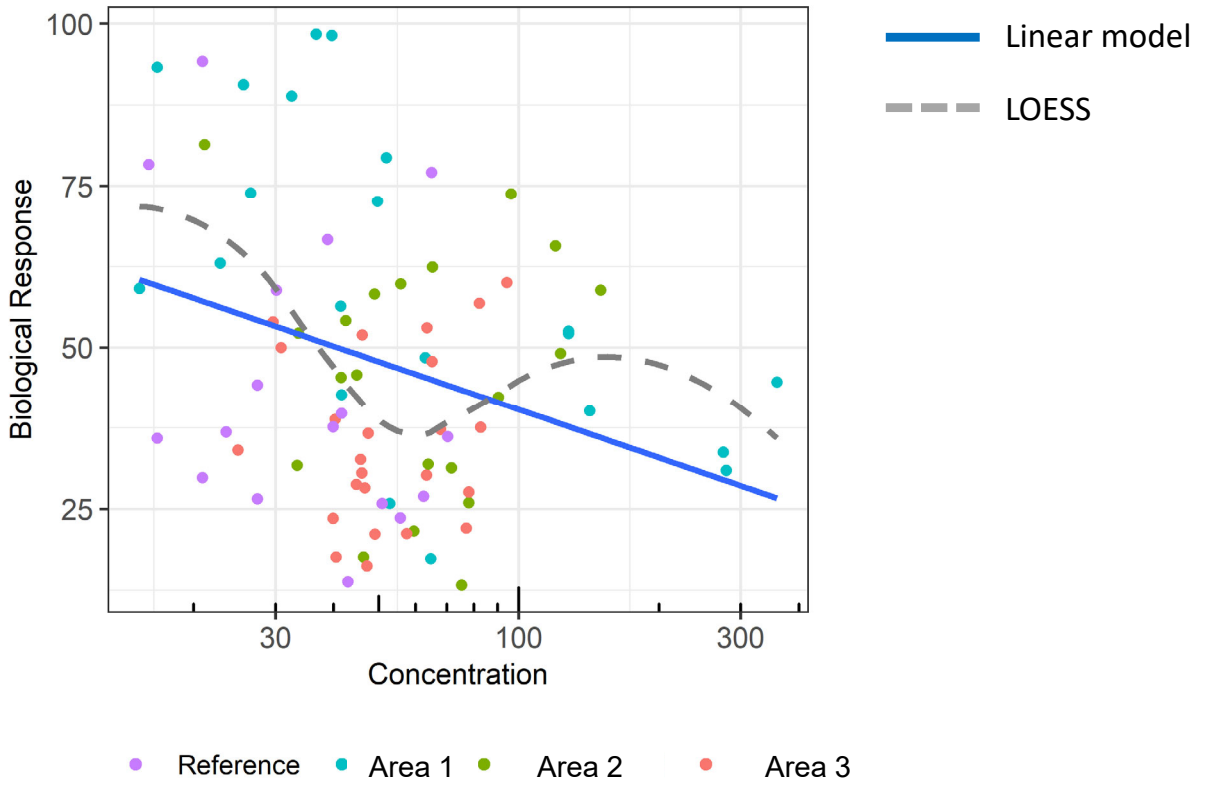
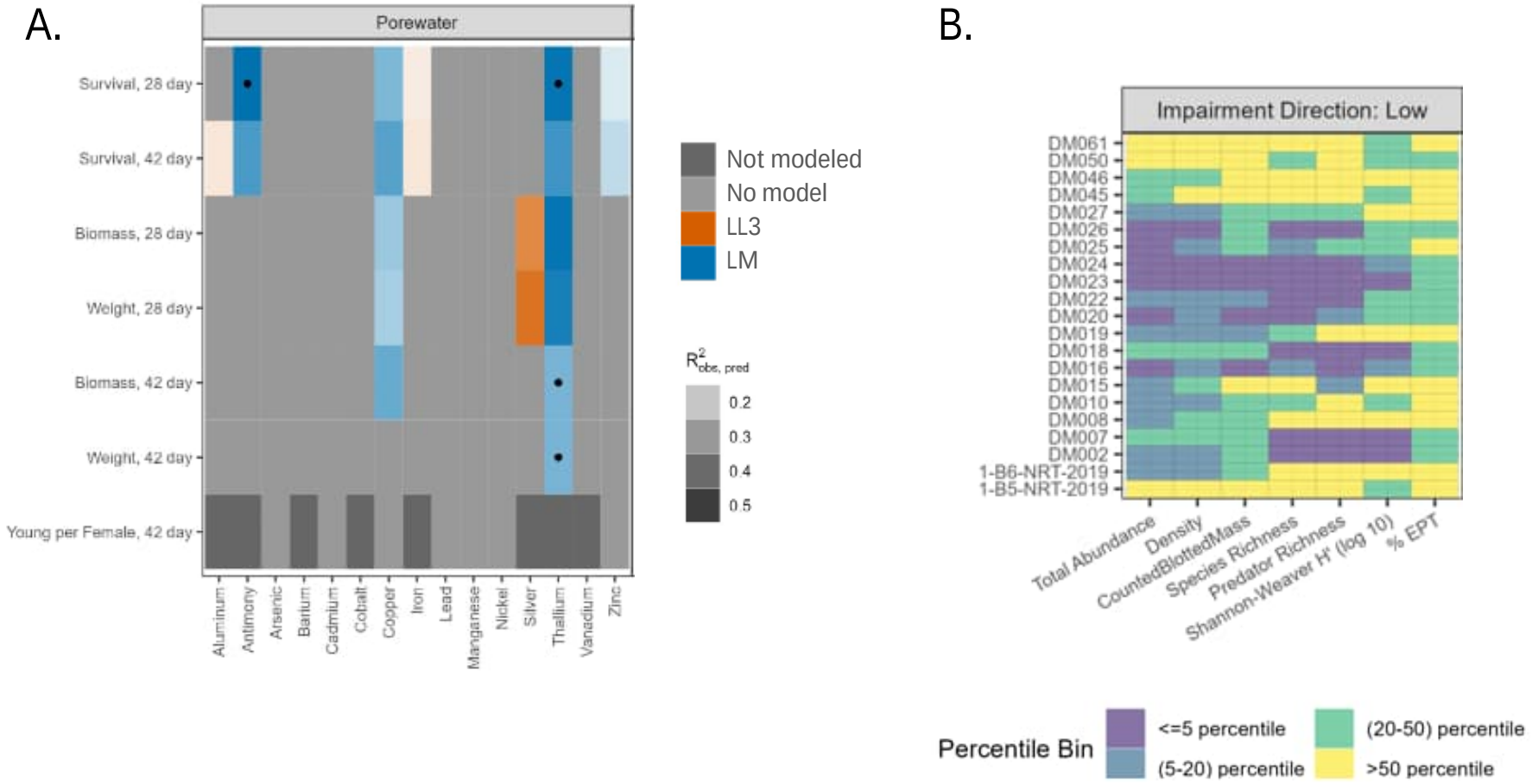



Figure L-2. Smoothing Functions on Scatterplots

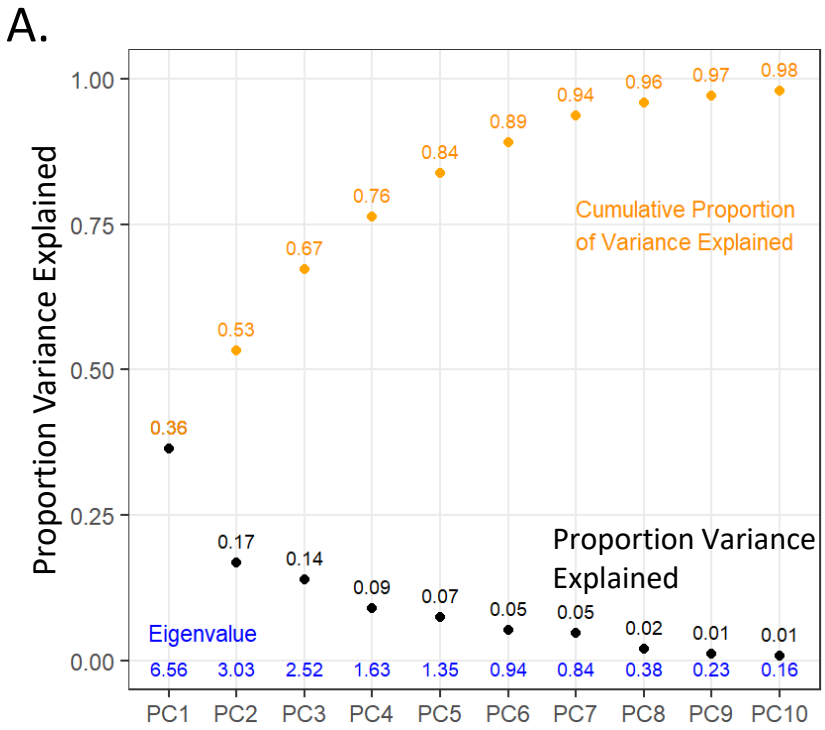
ERM




Notes:
(A) Mosaic showing the strength, direction, and significance of correlational tests. The intensity of the orange and blue colors corresponds to the observed $R^2$ of the given model type consistent with the gray-scale color ramp in the legend.
(B) Mosaic showing how benthic invertebrate community metrics in Site samples align with reference percentile bins.

Figure L-3. Mosaic Data Summaries

ERM

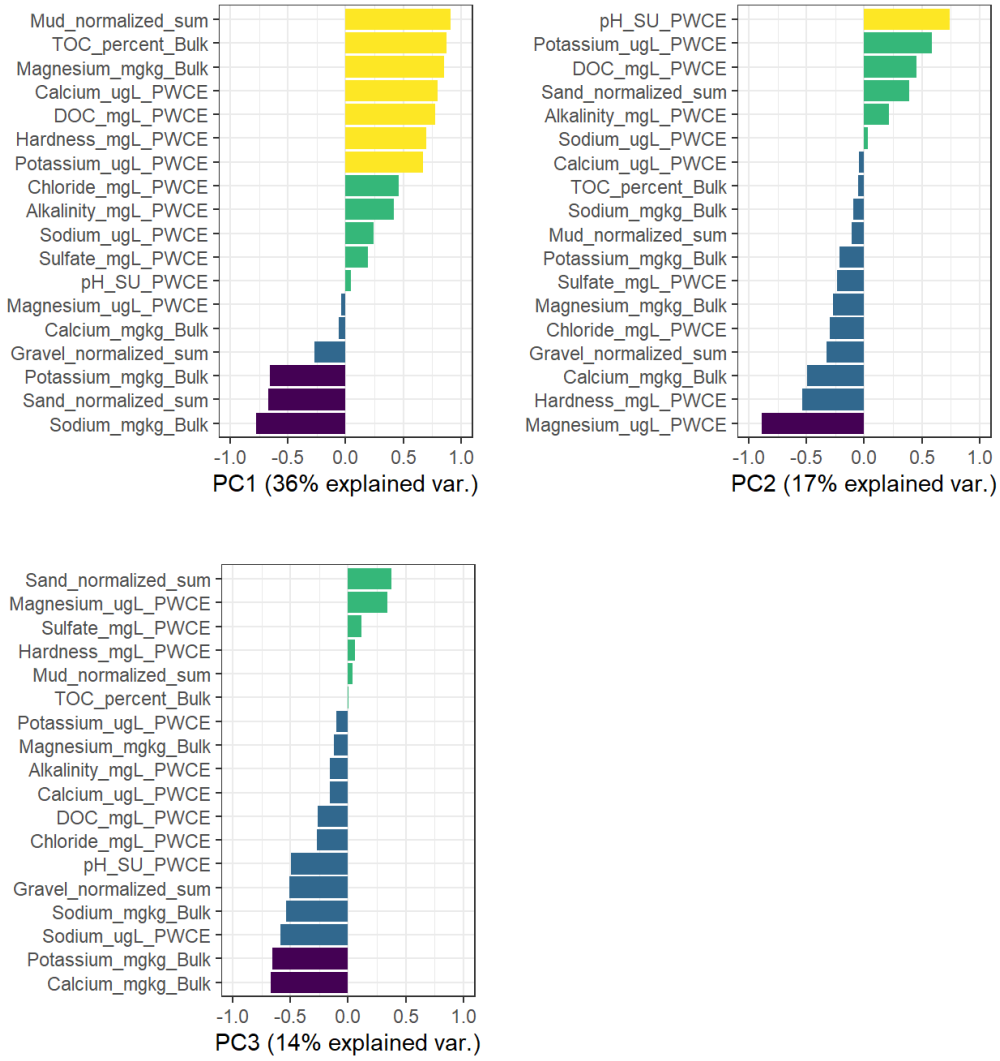



Notes:
All PCAs are reported with outputs similar to those in this figure series. (A) Scree plots show the proportion of variance explained with each PC axis and help determine the number of PCs that should be incorporated into future analyses.

Figure L-4a. PCA Outputs (Scree Plots)

ERM

B.

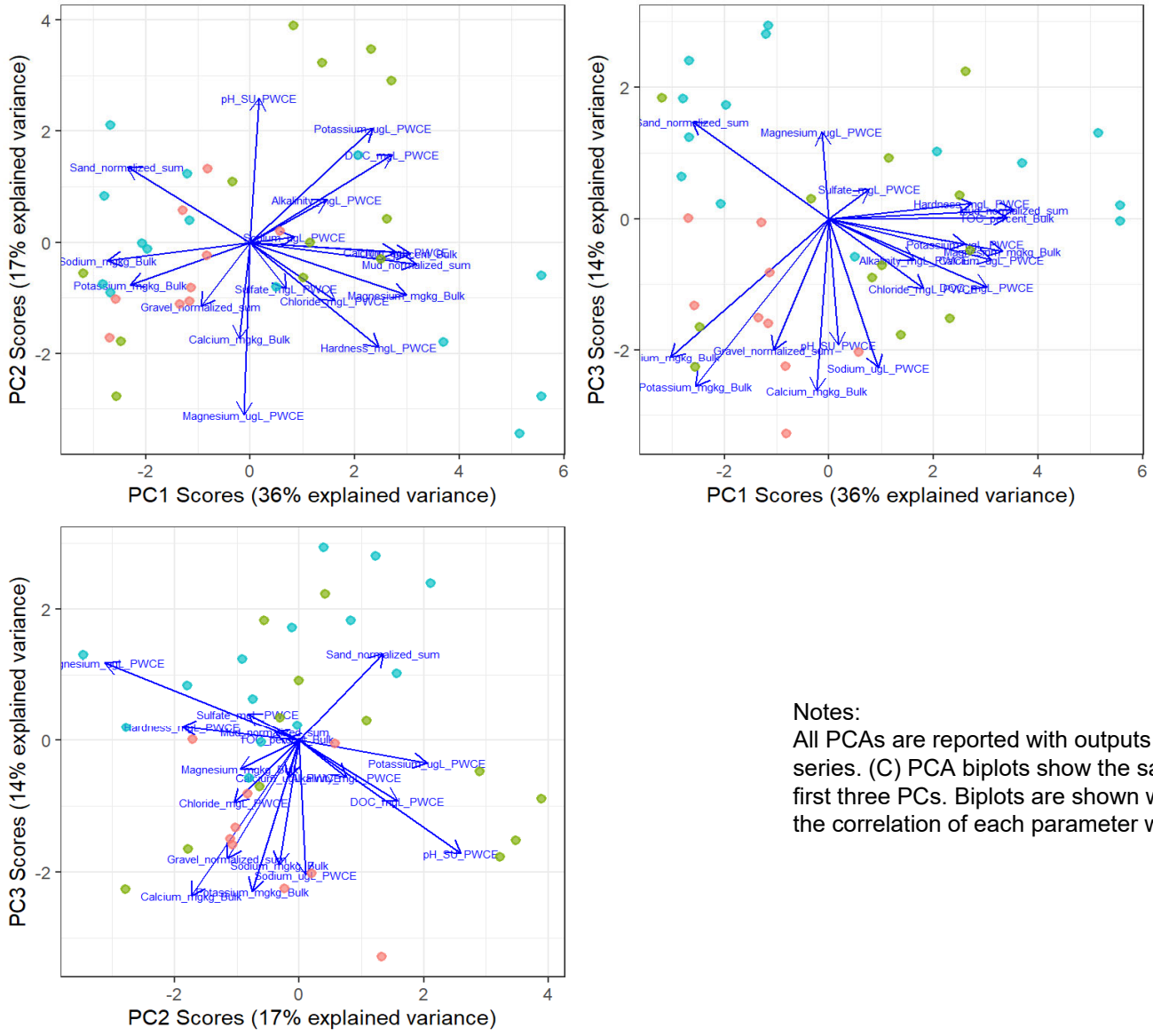



Notes:
All PCAs are reported with outputs similar to those in this figure series. (B) Loadings plots show the degree of correlation of each parameter with each PC. Color coding is used to indicate the strength of correlation in the positive or negative direction.

Figure L-4b. PCA Outputs (PC Loading Plots)

ERM

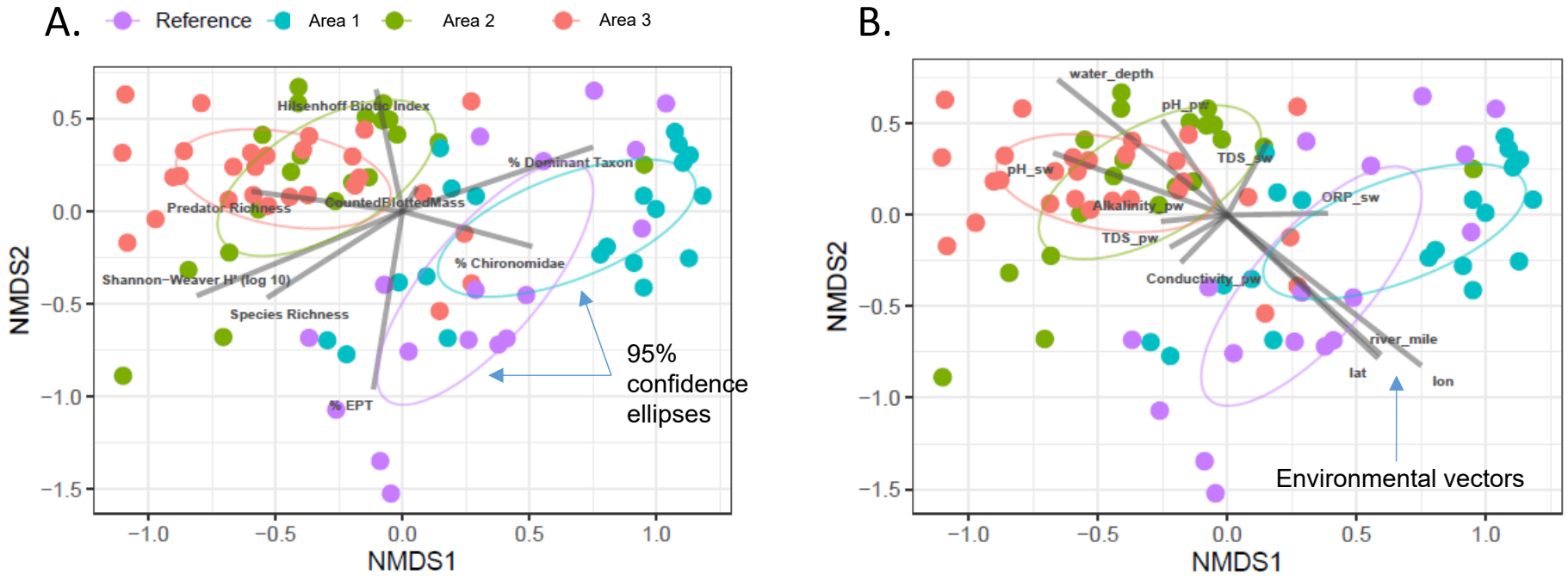
C.




Notes:
All PCAs are reported with outputs similar to those in this figure series. (C) PCA biplots show the sample scores on each of the first three PCs. Biplots are shown with vector overlays showing the correlation of each parameter with each PC.

Figure L-4c. PCA Outputs (PC Biplots)

ERM




Notes:
The first two dimensions of the NMDS output are displayed (i.e., NMDS 1 vs. NMDS 2). Figure A shows the vectors that denote correlation of abundance, tolerance and diversity metrics with the NMDS axes. Figure B shows the same NMDS output with a different set of vectors that denote correlation of environmental variables with the NMDS axes. For both figures, 95% confidence ellipses are constructed around samples from each of the four sampling areas in the study.

Figure L-5. Example NMDS Output

ERM

3-parameter sigmoid model (LL3)

$$y = \left(\frac{d}{1+exp^{(-b(\ln(x)-\ln(e))}}\right)$$

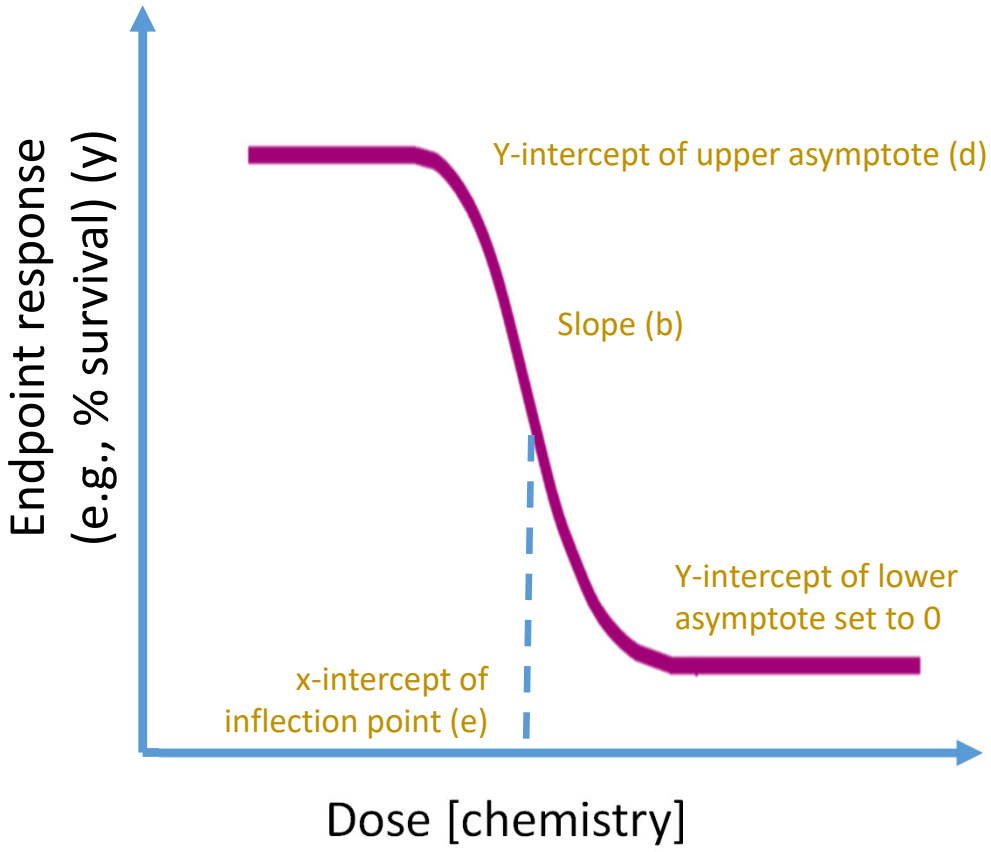



Figure L-6. Sigmoid Model Form (LL3)

ERM

Simple linear Regression

$$y = \mathrm{b} + m * log(x)$$

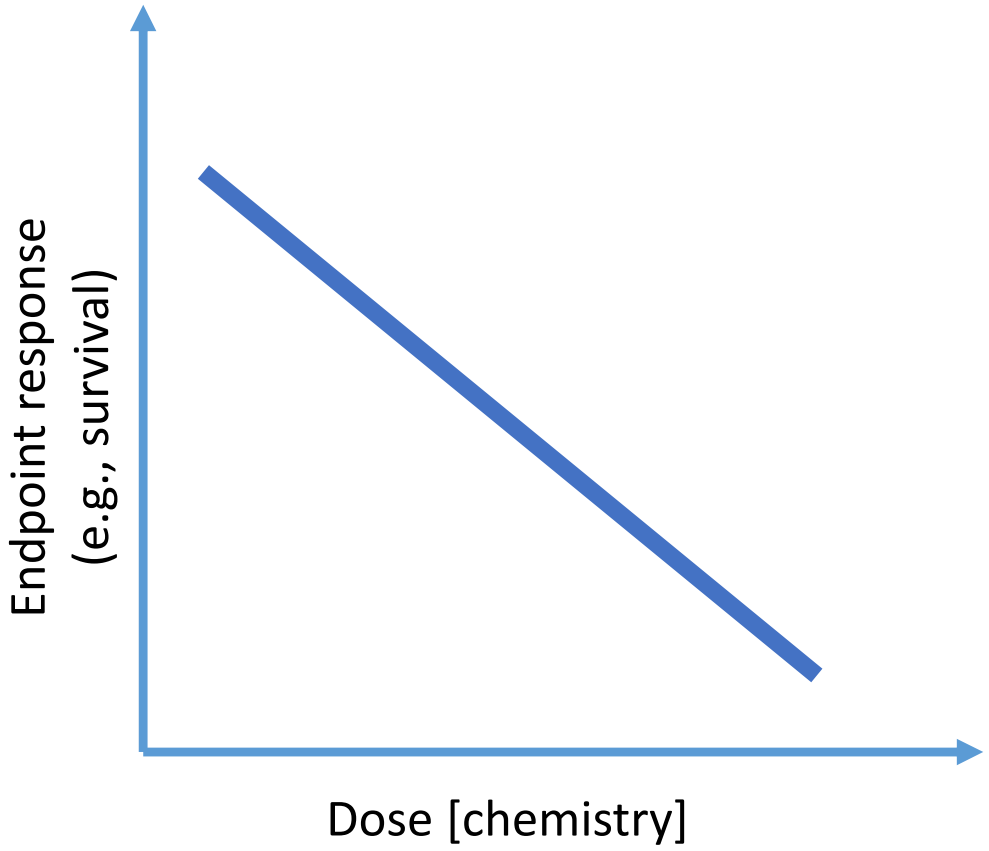
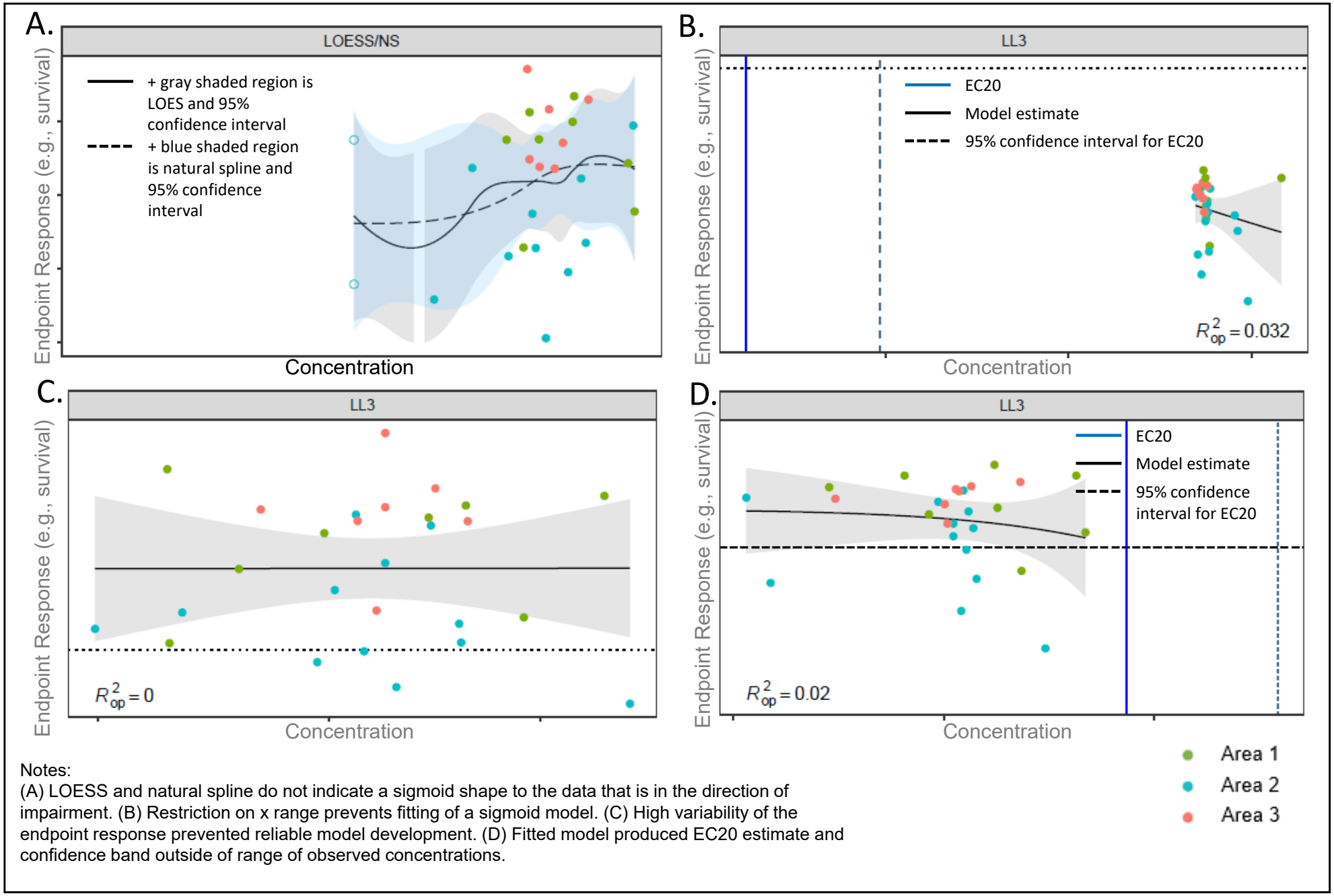



Figure L-7. Linear Model Form (LM)

ERM




Notes:
(A) LOESS and natural spline do not indicate a sigmoid shape to the data that is in the direction of impairment. (B) Restriction on x range prevents fitting of a sigmoid model. (C) High variability of the endpoint response prevented reliable model development. (D) Fitted model produced EC20 estimate and confidence band outside of range of observed concentrations.

Figure L-8. Examples of Data Sets Poorly Suited for Sigmoid Models

ERM

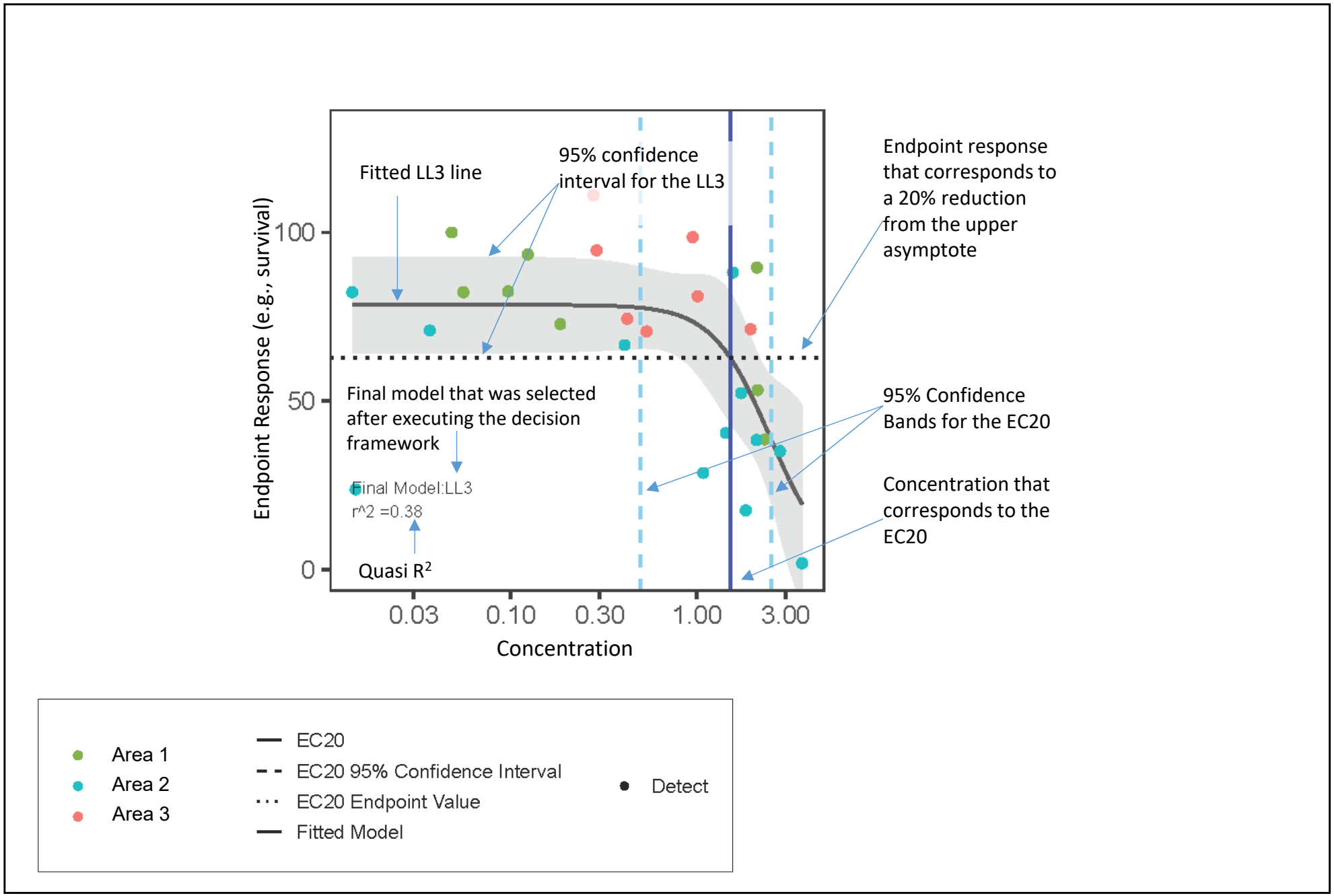



Figure L-9. Annotations for LL3 Bioassay Models

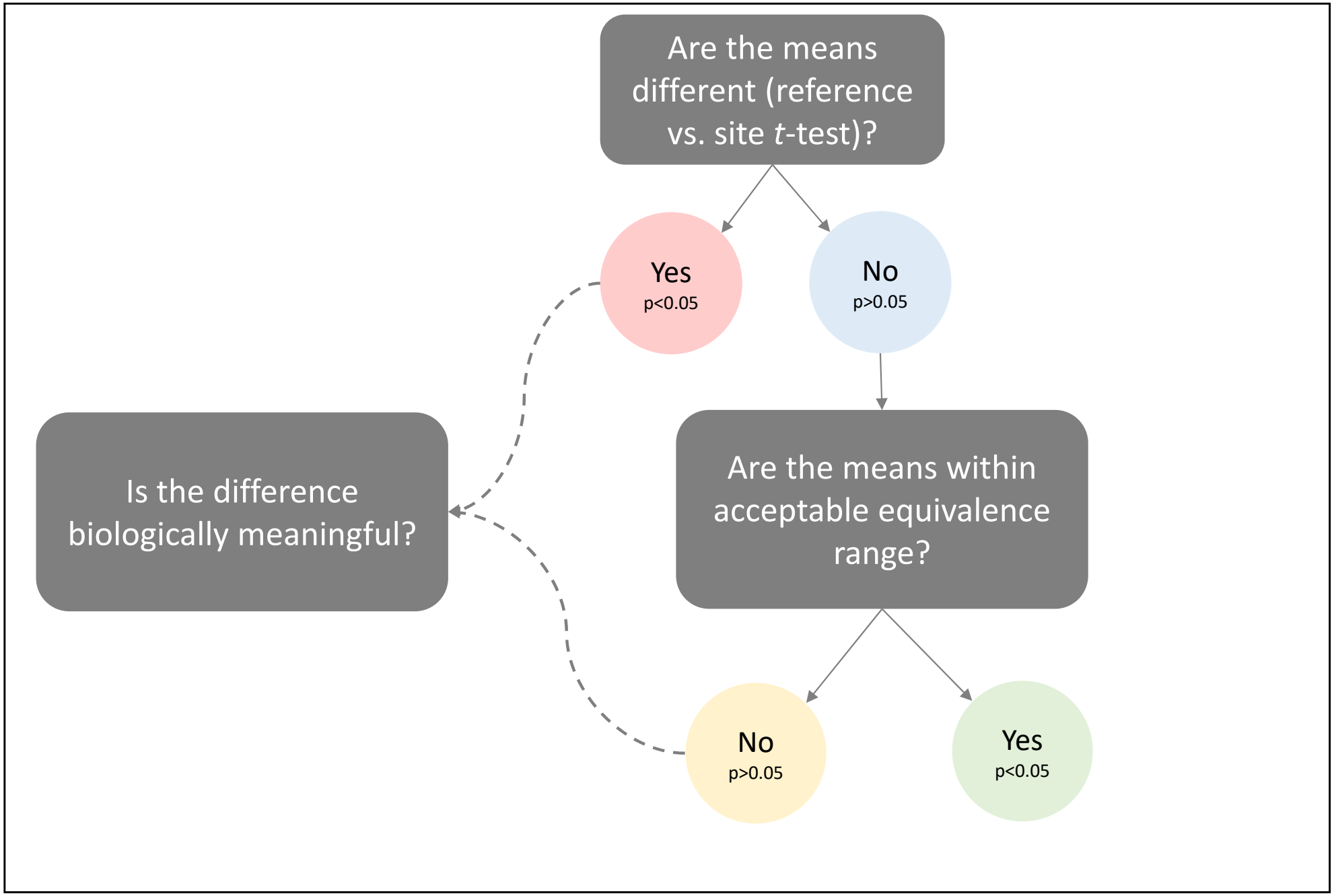
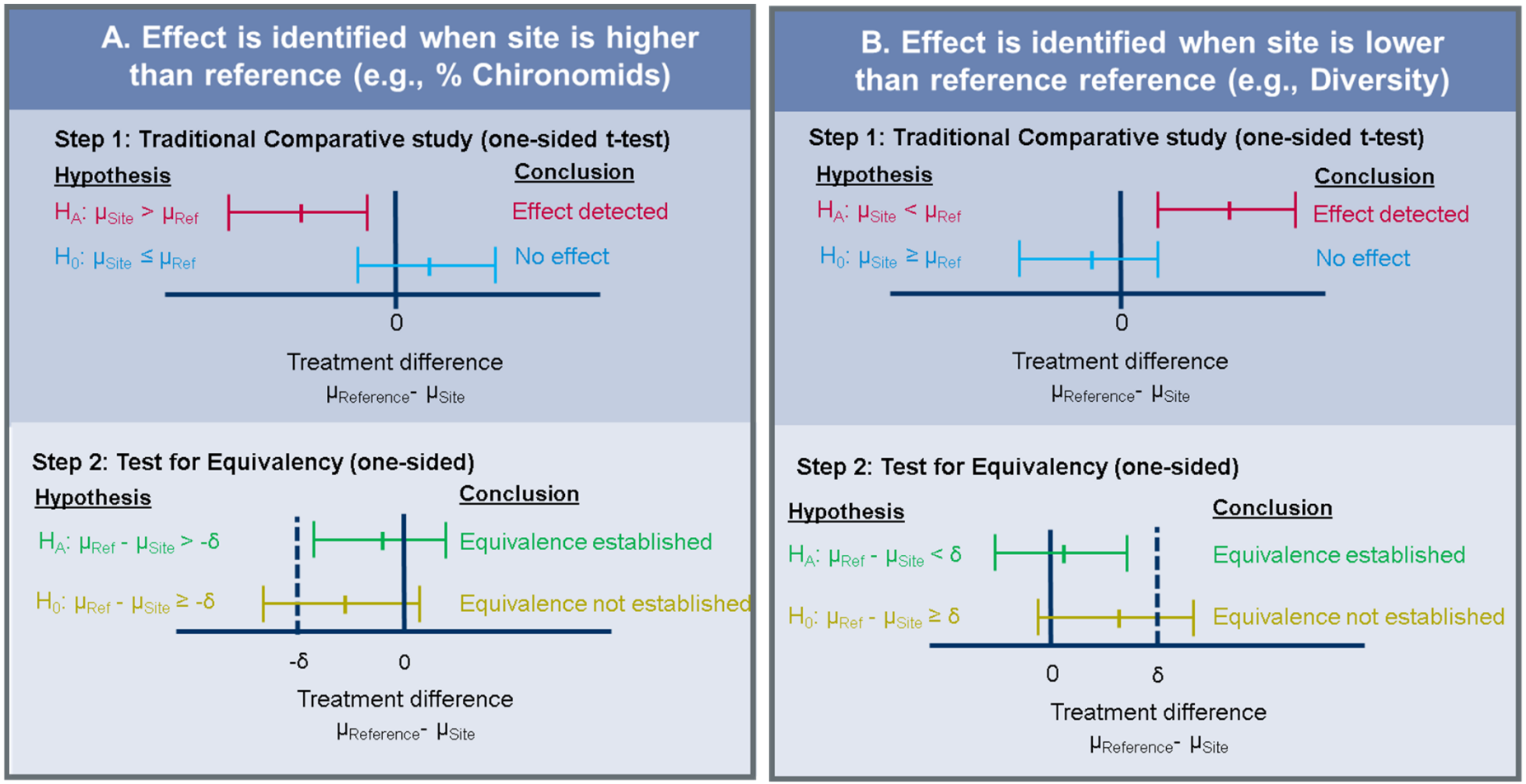



Figure L-10. Decision Tree for Equivalence Testing

ERM




Notes:
Schematics show the two-step process for determining whether the site is different from reference for (A) a metric where effects are identified if site is higher than reference and (B) a metric where effects are identified if site is lower than reference. The first step includes a traditional *t*-test. Where "no effect" is concluded in step 1, the analysis advances to step 2 to determine equivalence. Colored lines represent the mean and confidence interval. μ represents the population mean and δ represents the effect size that is established *a priori* as a difference that is meaningful for the study. Figure modeled after Walker and Nowacki (2010).

Figure L-11. Hypotheses for Equivalency Testing

ERM

ERM HAS OVER 140 OFFICES ACROSS THE FOLLOWING COUNTRIES AND TERRITORIES WORLDWIDE

Argentina
Australia
Belgium
Brazil
Brunei
Canada
Chile
China
Colombia
France
Germany
Guyana
Hong Kong
India
Indonesia
Ireland
Italy
Japan
Kenya
Malaysia
Mexico
Mozambique
Netherlands
Panama
Peru
Poland
Portugal
Romania
Singapore
South Africa
South Korea
Spain
Switzerland
Taiwan
Thailand
United Arab Emirates
United Kingdom
United States
Vietnam

**ERM's Carpinteria Office**
1180 Eugenia Place
Suite 204
Carpinteria, CA 93013

T: +1 805 684 2801
F: +1 805 684 1978

**www.erm.com**