APPENDIX D. DEVELOPMENT OF A BACKGROUND-TO-CRITERION REGRESSION MODEL

D.1. INTRODUCTION

Not all areas of the country have sufficient water chemistry and biological data to derive criteria for specific conductivity (SC) by the field method of calculating extirpation concentrations (XC_{95}) and hazardous concentrations (HC_{05} values) from an XC_{95} distribution (XCD) (see Section 3.1.3.). For such cases, the U.S. Environmental Protection Agency (EPA) is providing alternative methods that geographically extend results of the primary XCD method. This appendix describes a method to estimate criteria for new areas with different background SC using a background-to-criterion (B-C) model.

Background SC in a region and the associated HC_{05} are expected to be strongly related based on ecological and evolutionary theory and the observed responses of invertebrates to major ions (see Section 2.4 in the main document). The most salt-intolerant invertebrates occur in streams with the lowest background SC. As SC increases, the most salt-intolerant species are adversely affected and ultimately cannot persist. Thus, where regional background SC is higher, those taxa adapted to low SC are absent, and the SC level that is protective of 95% of taxa (HC_{05}) is higher.

EPA developed 24 XCDs using data from ecoregions with background SC ranging from 22 to 626 µS/cm. The prediction error rate of the B-C model was evaluated by a leave-one-out cross validation (LOOCV) (Arlot and Celisse, 2010).

Relatively salt-intolerant genera, as indicated by low XC_{95} values, occupy habitats in each region with the lowest ionic concentration. When both are log-scaled, the increase in background SC is linearly related to the HC_{05}. This regular and biologically relevant relationship between background SC and the HC_{05} confirms that the lower portion of the XCDs are similar in similarly exposed communities even though the represented genera may differ between ecoregions. The relationship between background SC and the HC_{05} identified from the XCD is sufficiently strong to identify a criterion continuous concentration (CCC) for areas with sufficient stream chemistry data but little or no paired biological data within an ecoregion or for new ecoregions.

Examples of B-C-modeled HC_{05} values for SC are provided for 62 Level III ecoregions (see Tables D-3−D-5). Using additional data sources, states are encouraged to assess whether the background value used to estimate an example ecoregional HC_{05} is reliable and representative and whether the ionic mixture of waters with elevated SC in the area of interest is related to sulfate and/or bicarbonate salts. Examples of B-C-modeled HC_{05} values were not calculated for ecoregions where the ionic pollutant was likely to be dominated by chloride, where the 25th centile background SC exceeded the B-C model parameters (>626 µS/cm), or when the sample size of the EPA-survey data was <25.

D.2. METHODS

Development of the B-C regression model required suitable data sets, which allowed calculation of XC_{95} and HC_{05} values for each ecoregion, estimation of background SC for each ecoregion, and regression of the HC_{05} values against the corresponding background SC values. There are 85 Level III ecoregions in the contiguous United States (Omernik, 1995, 1987) characterized by geology, physiography, vegetation, animal life, climate, soils, water quality, and hydrology.

D.2.1. Data Sets

For the purposes of this model development and validation, data requirements were relaxed relative to those for calculating a HC_{05} using the XCD method (i.e., fewer than 90 genera across 500 sites) (see Section 3.1.1.2).

State data sets from 48 ecoregions were considered. Five requirements were included:

- A minimum of 200 sampled sites for both biology and SC. This sample size is less than what is used to derive a stand-alone region-specific criterion. However, it is a sufficient sample size to derive HC_{05} values that are used as a group to represent a general relationship between the occurrence of salt-intolerant genera and regional SC levels represented in the data set.
- Taxonomic identification to genus of all individuals or a minimum sample or subsample of 100 individuals;
- Information about the ionic composition of streams in the region indicating that on a mass basis ($[HCO_3^-] + [SO_4^{2-}]) > [Cl^-]$, and $([Ca^{2+}] + [Mg^{2+}]) > ([Na^-] + [K^-])$;
- At least some samples with SC greater than 1,000 µS/cm to ensure a range of exposures; and

- More than 5% of XC_{95} values have 95% confidence bounds on the genus generalized additive model that intersect zero occurrence (see Section 3.1.2.1); that is, extirpation occurs in the sampled SC range.

Of the 48 data sets, 24 met all the requirements for estimating XC_{95} values for constructing the B-C model (see Table D-1). Two of the data sets were from distal ends of the Southeastern Coastal Plains (Ecoregion 65), Maryland and Mississippi, which are separated by more than 1,000 km. Background SC was different in the two data sets, so both were included for a total of 23 Level III ecoregions from 24 state data sets. The sources for the selected state data sets are described in Table D-1. Each selected state data set had been quality assured as a part of EPA's regionally approved monitoring program to assess water bodies as required by Section 305(b) of the Clean Water Act. Any site with a pH <6 was removed prior to analysis. The state data sets may represent either minimally affected or least disturbed (Stoddard, 2006) background SC for an ecoregion. In this document, natural background SC is defined as the range of ionic concentrations naturally occurring in waters that have not been substantially influenced by human activity. Minimally affected background is defined as the physical, chemical, and biological habitat found in the absence of significant human disturbance. Least disturbed background refers to the best available physical, chemical, and biological habitat conditions given the present state of the landscape. In the B-C model, background SC is the independent variable in the relationship between the SC experienced by a community of organisms in an area and the tolerance to SC of those sampled genera. The model may be updated as additional data become available.

Table D-1. State data sets used to develop the background-to-criterion (B-C) model

Level III ecoregion		Data source	Number of	
Number	Name	State data set[a]	Samples	Sites
15	Northern Rockies	Idaho DEQ	614	289
16	Idaho Batholith	Idaho DEQ	1,040	550
17	Middle Rockies	Idaho DEQ	510	306
19	Wasatch and Uinta Mountains	Utah DEQ	773	152
23	Arizona/New Mexico Mountains	Arizona DEQ	374	106
45	Piedmont	North Carolina DENR	665	433
47	Western Corn Belt Plains	Minnesota PCA	473	404
50	Northern Lakes and Forests	Minnesota PCA	734	596
51	North Central Hardwood Forests	Minnesota PCA	583	437
52	Driftless Area	Minnesota PCA	344	277
54	Central Corn Belt Plains	Illinois EPA	465	337
58	Northeastern Highlands	New York State DEC	383	383
59	Northeastern Coastal Zone	New York State DEC	277	277
60	Northern Allegheny Plateau	New York State DEC	562	562
64	Northern Piedmont	Maryland DNR	539	438
65	Southeastern Plains	Maryland DNR	359	289
65	Southeastern Plains	Mississippi DEP	457	361
66	Blue Ridge	North Carolina DENR	322	224
67	Ridge and Valley	West Virginia DEP	926	752
69	Central Appalachians	West Virginia DEP	1,661	1,420
70	Western Allegheny Plateau	West Virginia DEP	2,075	1,695
71	Interior Plateau	Indiana DEM	336	290
72	Interior River Valleys and Hills	Illinois EPA	460	340
83	Eastern Great Lakes Lowlands	NY	591	591

[a]DEC = Department of Environmental Conservation; DEM = Department of Environmental Management; DENR = Department of Environment and Natural Resources; DEP = Department of Environmental Protection; DEQ = Department of Environmental Quality; DNR = Department of Natural Resources; PCA = Pollution Control Agency.

D.3. ESTIMATING THE 95% EXTIRPATION CONCENTRATION (XC_{95}) AND CONSTRUCTING GENUS EXTIRPATION CONCENTRATION DISTRIBUTIONS (XCDS)

The process for developing the B-C model is shown in Figure D-1 (see Section 3.1.3). Prior to developing of the B-C model, a HC_{05} value was derived for each of the 24 data sets. The HC_{05} is the 5th centile of the genera XC_{95} values represented in the XCD. The XC_{95} values were calculated for genera occurring in >25 samples. Occurrences of a genus were weighted to adjust for uneven sampling throughout the sampled range of SC in each data set. Sets of XC_{95} values were ranked from lowest to highest SC, and the HC_{05} was estimated by interpolation at the 5th centile for each data set. More than 5% of the XC_{95} values were unambiguously defined within the tested range (see Table D-2) based on an inspection of scatter plots and generalized additive models of the probability of observing each genus in each data set (see Section 3.1.2.1). Paired 25th centile SC and SC HC_{05} values used to produce the B-C model are listed in Table D-2.

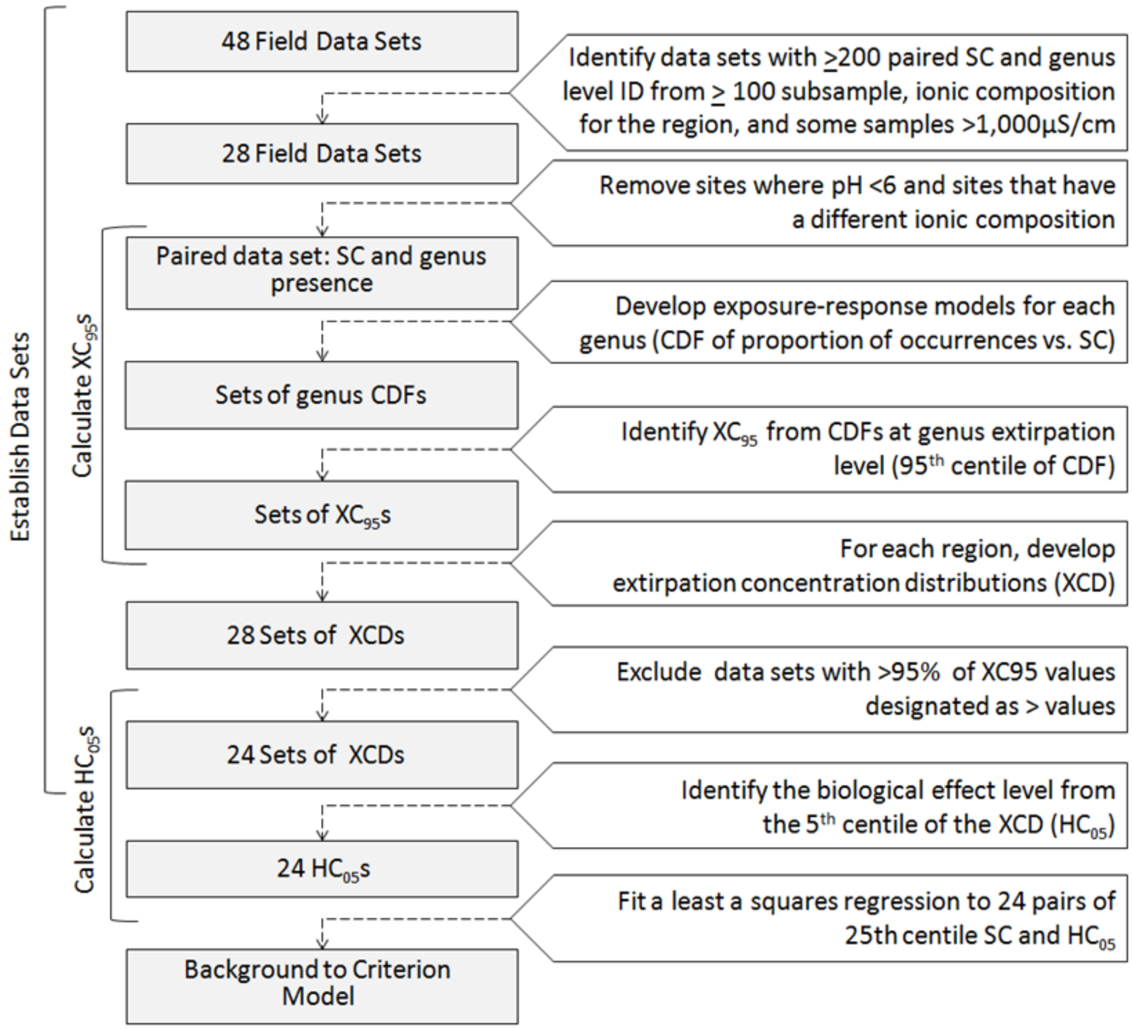

Figure D-1. Process for developing background to criterion least squares regression model.

Table D-2. Paired biological and specific conductivity data used to produce the background-to-criterion (B-C) model.

Level III ecoregion	State data set[a]	Number of genera	Proportion of XC_{95} values > observed range	25th centile SC, state data sets (µS/cm)	B-C HC_{05}[b] (µS/cm) for model
15	Idaho DEQ	120	0.75	29	142
16	Idaho DEQ	136	0.79	42	185
17	Idaho DEQ	110	0.90	136	264
19	Utah DEQ	69	0.81	271	299
23	Arizona DEQ	97	0.84	191	249
45	North Carolina DENR	212	0.74	68	138
47	Minnesota PCA	106	0.89	587	934
50	Minnesota PCA	176	0.87	108	320
51	Minnesota PCA	147	0.78	325	494
52	Minnesota PCA	83	0.88	534	655
54	Illinois EPA	124	0.54	626	813
58	New York State DEC	72	0.90	50	212
59	New York State DEC	42	0.86	350	706
60	New York State DEC	72	0.88	132	248
64	Maryland DNR	89	0.52	150	257
65-N	Maryland DNR	92	0.43	103	257
65-S	Mississippi DEP	127	0.73	38	131
66	North Carolina DENR	176	0.65	22	69
67	West Virginia DEP	123	0.82	59	154
69	West Virginia DEP	142	0.73	94	305
70	West Virginia DEP	139	0.63	169	338
71	Indiana DEM	69	0.77	296	479
72	Illinois EPA	116	0.91	460	1,108
83	New York State DEC	68	0.81	272	525

[a]The data requirements of the full method using a minimum of 500 paired biological and chemical samples were not met in all ecoregions; therefore, the calculated HC_{05} values are not necessarily robust enough for example criteria but are reliable enough for development of the B-C model based on the predictive performance of the model.
[b]HC_{05} = hazardous concentration of the 5th centile of a taxonomic extirpation concentration distribution.

D.3.1. Background-to-Criterion Development and Validation

The B-C model was developed using paired background SC and HC_{05} values, listed in Table D-2, that were calculated as described for the XCD method (see Section 3.1). HC_{05} values were calculated from XCDs from the 24 data sets. Background SC was estimated at the 25th centile from these same data sets. A linear least squares regression was used to develop the B-C model using the log of background SC as the independent variable and the log of HC_{05} as the dependent variable. The 50% prediction interval (PI) was also calculated.

Model validation used a LOOCV procedure (Arlot and Celisse, 2010; James et al., 2013). This involves removing one ecoregional data pair at a time and recalculating the B-C model with the remaining 23 ecoregional pairs to produce a test model (see Figure D-2). A predicted HC_{05} is calculated for the observed background SC from the removed data pair. The observed HC_{05} is subtracted from the predicted HC_{05} and the difference is squared to yield a squared error (SE) for the B-C model validation. This is repeated for all 24 state data sets to generate 24 SEs and the root mean squared error (RMSE) and coefficients of variation (CV) are calculated for all 24 SEs values. The measurement of fit of the B-C model is summarized by the validation RMSE and its coefficient of variation. A small difference between the validation RMSE and the B-C model's RMSE indicates a good model. A small CV indicates little variation in the SEs.

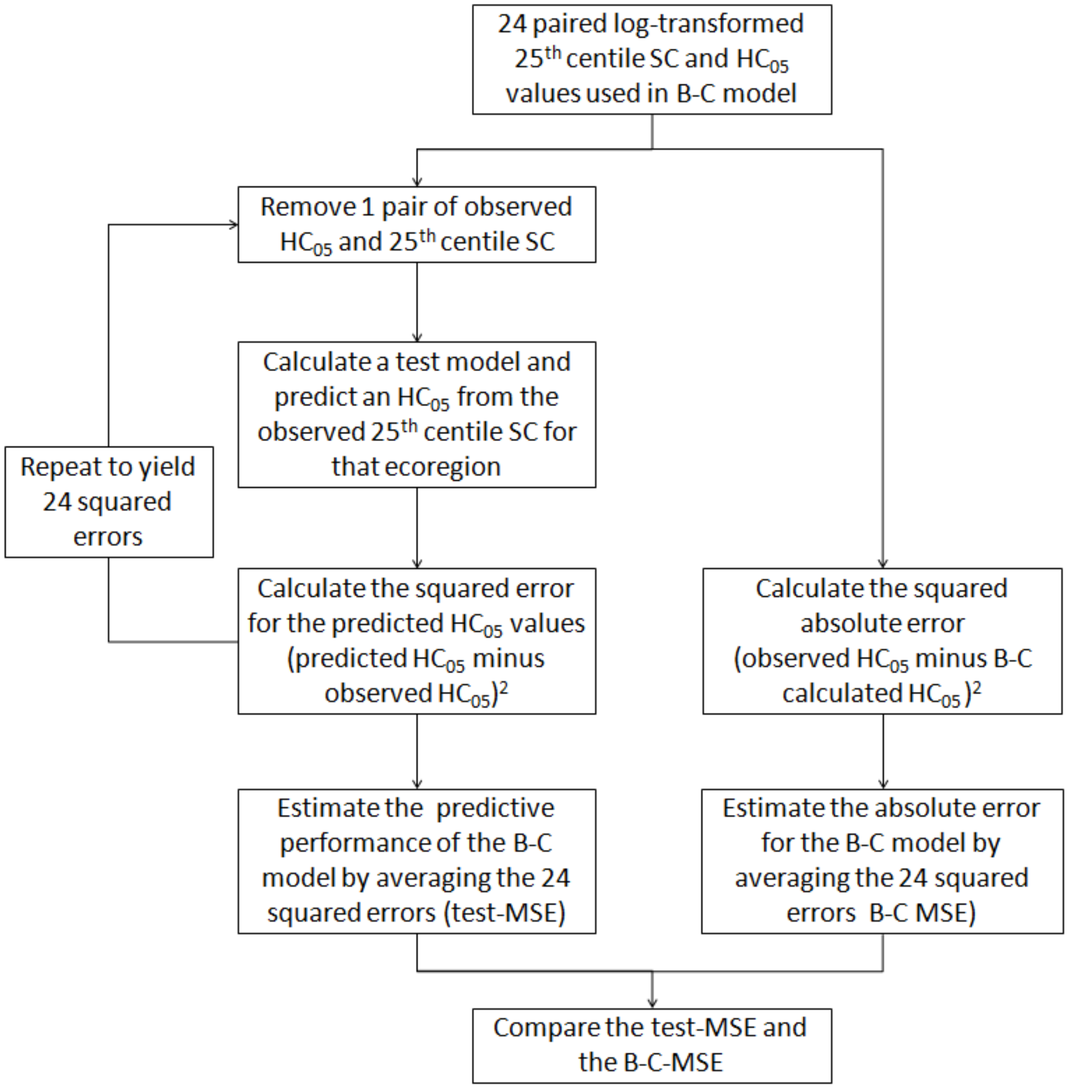

Figure D-2. Process for validating background to criterion least squares regression model.

D.4. RESULTS

D.4.1. Level III Ecoregional Specific Conductivity Background

Background SC was estimated from 24 state data sets from 23 Level III ecoregions. The 25^{th} centile values for the 24 data sets ranged from 22 µS/cm in Ecoregion 66 to 626 µS/cm in Ecoregion 72 (see Table D-2).

The XCDs for the 24 data sets are shown in Figure D-3. Only the lower 30% is shown because many of the XC_{95} values in the upper portion of the XCDs are greater than the calculated value and are not defined except in a relative sense. Plotting the lower 30% more clearly shows the region of the XCD containing the HC_{05}.

As the background SC increases, the XCDs for the ecoregions plot further to the right at higher SC (see Figure D-3). Similarly, the HC_{05} identified from the intercept of each XCD with the 5th centile line, increases as background SC increases. The genera that contribute most directly to each HC_{05} are the most salt-intolerant genera in each ecoregion.

The HC_{05} values were calculated at the 5th centile of the 24 XCDs shown in Figure D-3. Having derived HC_{05} values and estimates of measured background SC for 24 data sets (see Table D-2), the log10-transformed HC_{05} values were regressed against log10-transformed background SC values to generate a predictive model (see Figure D-4). Each circular point on the graph represents the relationship between background SC for a Level III ecoregion and the HC_{05} for that ecoregion.

The regression model using the estimated background SCs from the 24 state data sets yielded a strong model ($r = 0.93$). The lower 50% prediction limit (PL) from the mean regression line identified a 75% probability that a HC_{05} derived by the XCD method for a new ecoregion would be equal to or greater than the *y*-coordinate of the lower prediction limit. The PLs were used to determine whether an XCD calculated HC_{05}, or the mean or lower PL was the more reasonable SC choice for the HC_{05} (see Sections 3.7.2 and D.5).

The B-C model represents a range of minimally affected or least disturbed SC background conditions (22−626 µS/cm) and HC_{05} values (69 to 1,108 µS/cm). The B-C model is described by eq D-1:

$$Y = 0.657\,X + 1.075 \qquad \text{(D-1)}$$

Where:

X is the log10 of the ecoregion background SC (µS/cm).

Y is the log10 of the predicted HC_{05} (µS/cm).

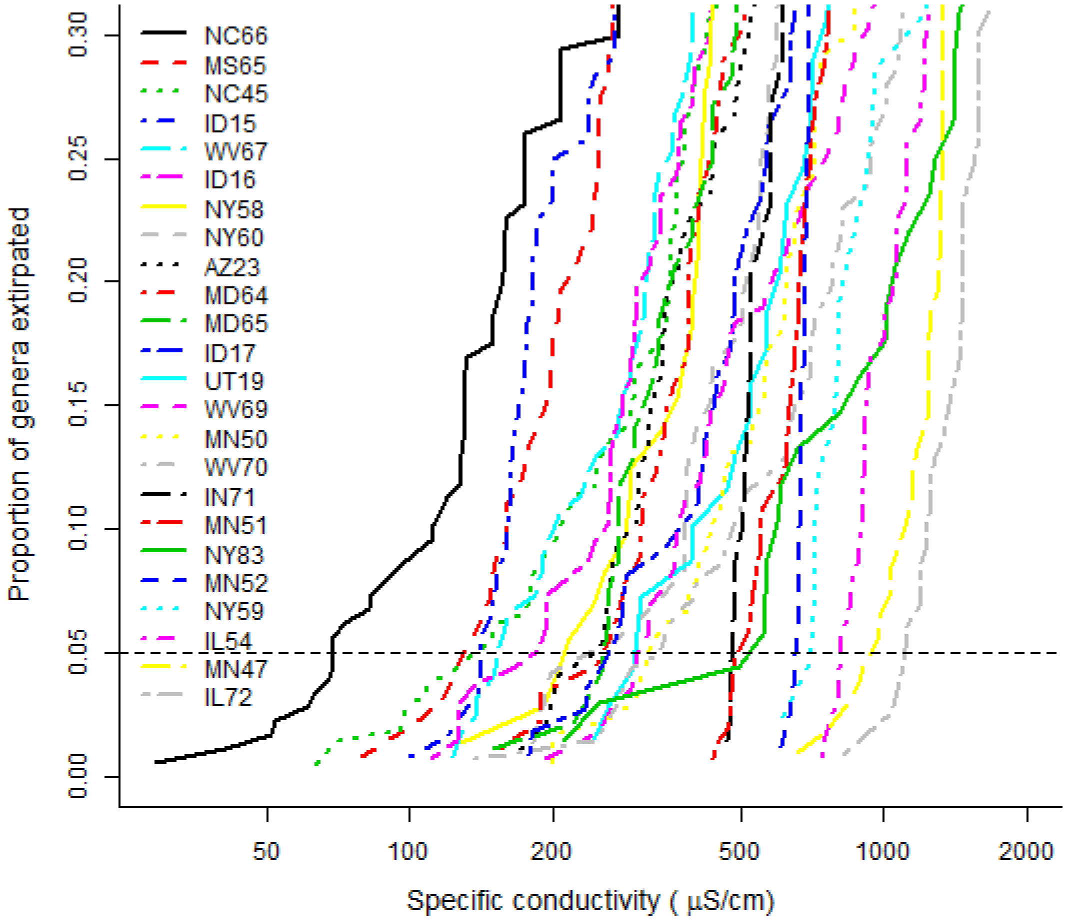

Figure D-3. **The lower 30% of the 24 extirpation concentration distributions (XCDs).** Hazardous concentration (HC_{05}) is the specific conductivity (SC) at the 5^{th} centile of each XCD (horizontal dashed line). The ecoregion legend and XCDs are ordered from lowest (NC66) to highest (IL72) HC_{05} and roughly from lowest to highest background SC. Untransformed SC values shown on log10 scaled *x*-axis.

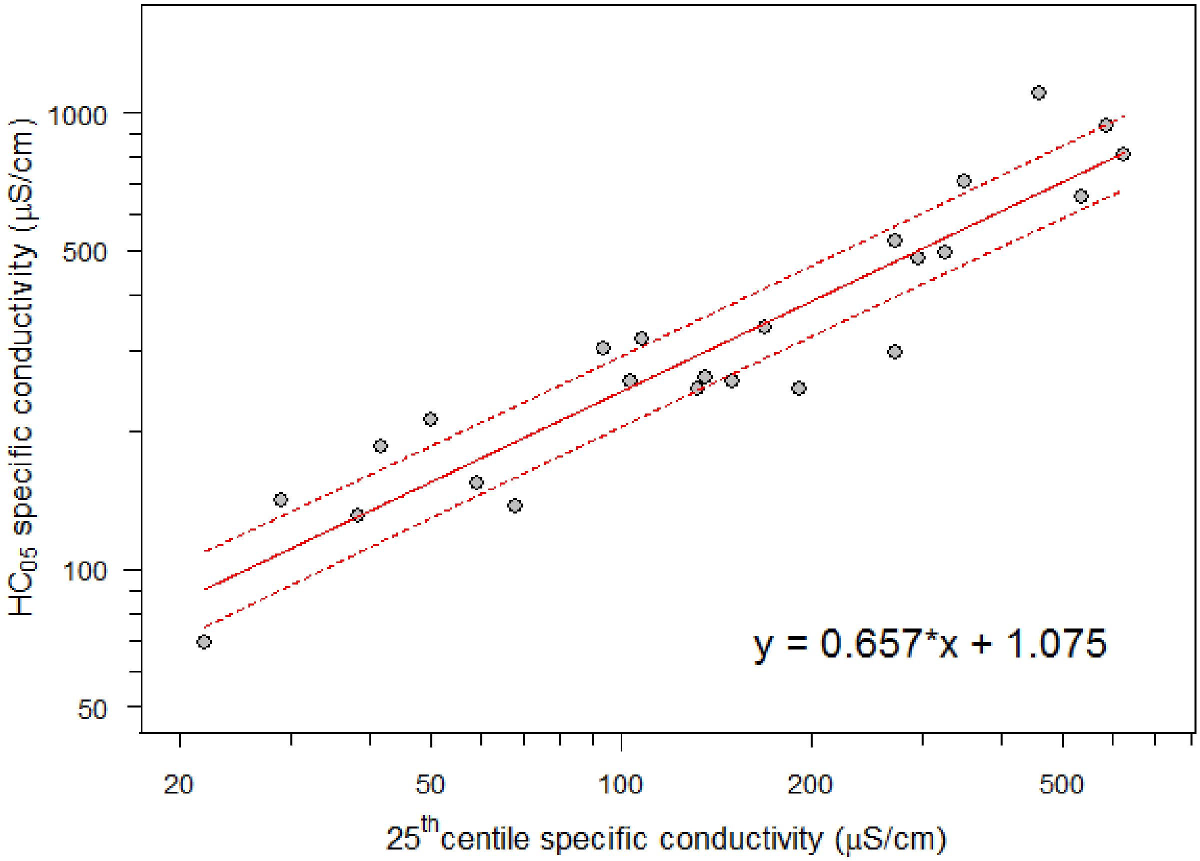

Figure D-4. **Benchmark-to-criterion empirical model.** The *x*-axis shows the untransformed background specific conductivity (SC) estimated at the 25th centile for 24 state data sets, scale is log10. Also scaled (log10), the *y*-axis shows the untransformed 5th centile hazardous concentrations (HC_{05}) of extirpation concentration distributions (XCD) shown in Figure D-3. Solid line is the log10-log10 linear regression line. Dotted lines demarcate the 50% prediction intervals (PI). That is, the probability that any new HC_{05} would plot within those bounds is 50% and only 25% are expected to fall below the lower prediction limit (PL). Pearson's correlation coefficient ($r = 0.93$). Data in Table D-2.

The B-C model was validated using a LOOCV procedure. Pairs of observed and predicted values are shown in Figure D-5. The log10 residual sums of squares (overall root mean square) for the cross validated model is 0.115, which is a corrected measure of prediction error, averaged across all 24 regions, and its CV is 0.046. The log10 RMSE for the B-C model derived from all 24 XCD is 0.105 µS/cm, and its CV is 0.042. Both RMSEs and their CVs are small and the small difference between the LOOCV RMSE and the absolute RMSE indicates that the model is supported by this validation exercise.

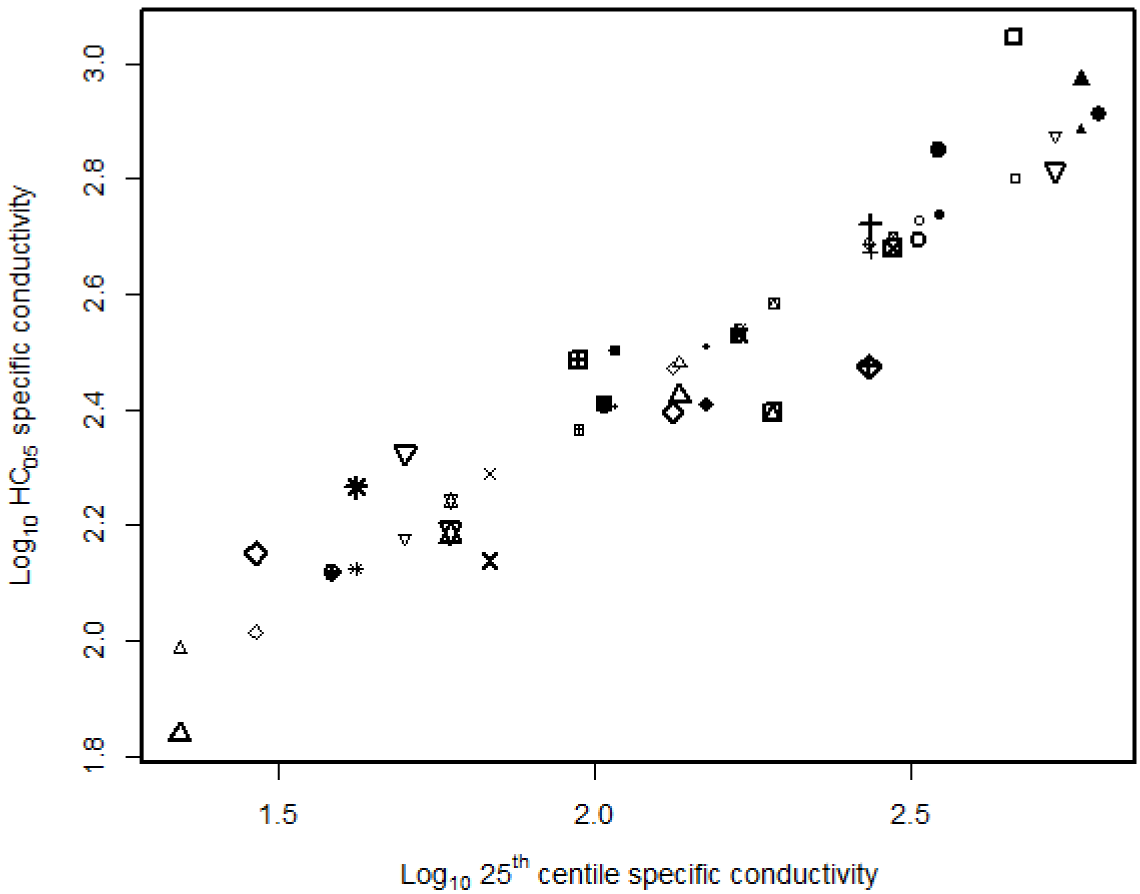

Figure D-5. Evaluation of the error in the background-to-criterion (B-C) model by comparison to values for 24 leave-one-out regression models. A different symbol is used for each ecoregion. The large symbols are the observed background versus hazardous concentration (HC_{05}) values from 24 data sets in the B-C model, as in Figure D-4. Each corresponding small symbol represents the predicted HC_{05} value for a particular region/data set estimated from a model that was built from a LOOCV set consisting of the other 23 data sets.

D.5. SUMMARY OF MODEL DEVELOPMENT

Within a relatively broad range of background SC (22 to 626 µS/cm), 24 XCDs were developed using data from 23 Level III ecoregions. Genera occupied all habitats including those with low XC_{95} values. This indicates that although the representative taxa may differ, there are genera with low SC tolerances that occupy low SC waters.

EPA also developed a B-C model. When the background SC and HC_{05} for a region are log scaled, the increase in background SC is linearly related to HC_{05}. This regular and biologically relevant relationship confirms that the lower portion of the XCDs are similar in similarly exposed communities even though the representative genera may differ between ecoregions. The model was validated using a LOOCV procedure which showed that the predictive performance of the model was strong.

The relationship between background SC and the HC_{05} identified from the XCDs is sufficiently strong to identify a HC_{05} for areas without biological sampling within an ecoregion or for new ecoregions. This B-C regression model was developed using biological data paired with SC data from waters with ionic mixtures dominated by calcium, magnesium, sulfate and bicarbonate ions and where background SC did not exceed 626 µS/cm. Therefore, the model is only appropriate for waters with similar ionic characteristics. Where on a mass basis $([HCO_3^-] + [SO_4^{2-}]) < [Cl^-]$, and $([Ca^{2+}] + [Mg^{2+}]) < ([Na^-] + [K^-])$, the model has not been thoroughly tested and professional judgment is required. In particular, the B-C model is not appropriate for NaCl inputs such as produced fracking water backflow (Haluszczak et al., 2013, Entrekin et al., 2011; Gregory et al., 2011; Veil et al., 2004) or road salt (Forman and Alexander, 1998; Kelly et al., 2008; Environment Canada and Health Canada, 2001; Evans and Frick, 2001; Kaushal et al., 2005). However, in the absence of an independent model, mixtures dominated by sodium, sulfate and bicarbonate ions (e.g., produced water from coal bed methane production (Brinck et al., 2008; Dahm et al., 2011; Jackson and Reddy, 2007; National Research Council, 2010; Clark et al., 2001; Veil et al., 2004) may be defensible given that the toxicity of these mixtures are more similar to that of calcium, magnesium, sulfate and bicarbonate ions compared to NaCl (Mount et al., 2016; Kunz et al., 2013; Soucek and Dickinson, 2015).

The HC_{05} values calculated in the B-C model were further evaluated as potential CCC because some of the modeled HC_{05} values were estimated using small data sets with an upper SC range that limited the measurement of XC_{95} values. The approach described in Section 3.7.2.1

was used and is summarized in Section D.6. Using that decision process, EPA calculated 62 candidate HC_{05} values for Level III ecoregions in the conterminous United States with comments on relative confidence in those estimates (see Tables D-3, D-4, and D-5). HC_{05} values were not derived for Level III ecoregions with NaCl dominated matrices, estimated background SC outside the range of the model (>626 µS/cm), or for ecoregions lacking sufficient data to estimate a background SC (N <20) (see Table D-6).

D.6. CALCULATION OF HC_{05} VALUES

D.6.1. Estimating the Background Specific Conductivity

Background SC was estimated at the 25th centile of the available state and EPA survey data sets and evaluated as being representative of minimally affected or least disturbed background for the ecoregion. Background represented minimally affected when the 25th centile was less than the mean natural base-flow SC estimated from a geophysical model (Olson and Hawkins, 2012). Background was characterized as least disturbed when the 25th centile measured from the field was greater than the mean natural base-flow SC estimated from a geophysical model. The predicted mean natural base-flow SC was calculated using geology, climate, soil, vegetation, topography, and other factors calibrated with reference sites (Olson and Hawkins, 2012).

This base-flow model uses a random forest method to generate predictions for natural base-flow SC values in streams. The same suite of predictor variables that are used to develop the base-flow water chemistry model are then generated for each stream line within the National Hydrography Dataset Plus version 2 (NHDPlusV2) with algorithms and code from the StreamCat Dataset (Hill et al., 2016). The StreamCat Dataset and algorithms provide catchment summaries of landscape data for all 2.65 million streams within the NHDPlusV2. All StreamCat code and quality assurance procedures are documented online via the EPA's StreamCat website (https://www.epa.gov/national-aquatic-resource-surveys/streamcat) and the StreamCat GitHub repository (https://github.com/USEPA/StreamCat). Briefly, StreamCat used the ArcGIS (ESRI, 2012) zonal-statistics-as-table tool to compute summary statistics of geospatial layers (e.g., geology) for each NHDPlusV2 catchment, that is, the local landscape that flows to a single stream segment. For a complete description of the geospatial framework used to generate StreamCat, see Hill et al. (2016). These geospatial layers included rock chemistry and unconfined compressive strength developed by Olson and Hawkins (2012), as well as several climatic and soil variables. Once these local summaries are generated, the algorithm then uses open source Python code and topological (from-to) flow tables provided with the NHDPlusV2 to accumulate the summaries and generate full catchment summaries of these geospatial layers, such as mean soil bulk density or mean whole rock sulfur content within each watershed. Once these data were generated, the random forest model is applied to the tables to produce predictions of natural base-flow water chemistry for all 2.65 million streams within the NHDPlusV2. Predicted mean natural base-flow water chemistry for each stream is then linked to NHDPlusV2 shapefiles through a unique identifier and mapped for the conterminous United States. The mean stream length-weighted SC within EPA Level III ecoregions is used to predict mean natural base-flow SC at the Level III ecoregional (see Figure D-6). The data sources are listed as a table in this Appendix in D-7.

Tables D-3, D-4 and D-5 list the SC 25th centiles from the state data sets, the EPA survey data sets, and the predicted mean natural base flow for each of the 85 ecoregions.

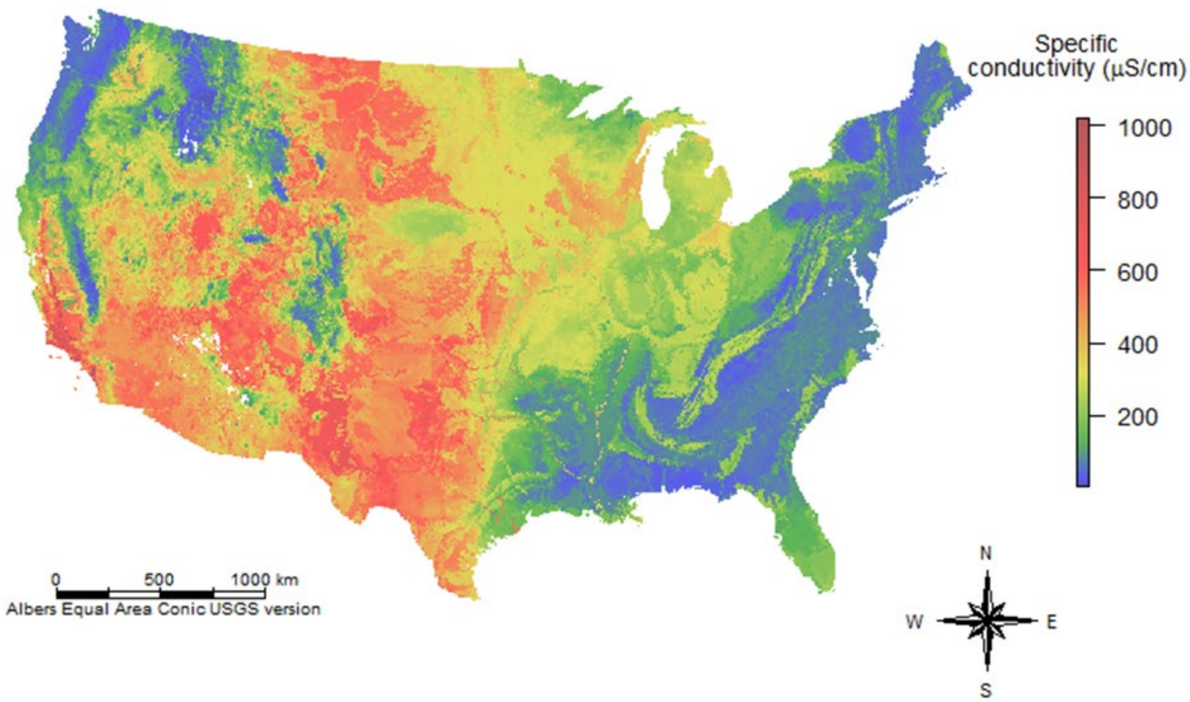

Figure D-6. Predicted natural base-flow for stream specific conductivity (SC). Mountainous regions tend to have lower background SC, the central plains and arid lands tend to have the highest, and other ecoregions have intermediate background SC.

D.6.2. Using Background to Calculate a 5th Centile Hazardous Concentration (HC_{05})

The HC_{05} for a defined geographic area or ecoregion without a sufficient data set or without suitable biological data is calculated using the background SC of that area or region in the B-C model. The decision tree for calculating and then choosing a suitable HC_{05} from background SC is shown in Figure D-7. Equation D-1 is used to calculate the mean HC_{05} and eq D-2 calculates the lower 50% PL for the area or region. Where suitable biological and water chemistry data are available, a HC_{05} can be calculated using formulae described in (see Section 3.1.3).

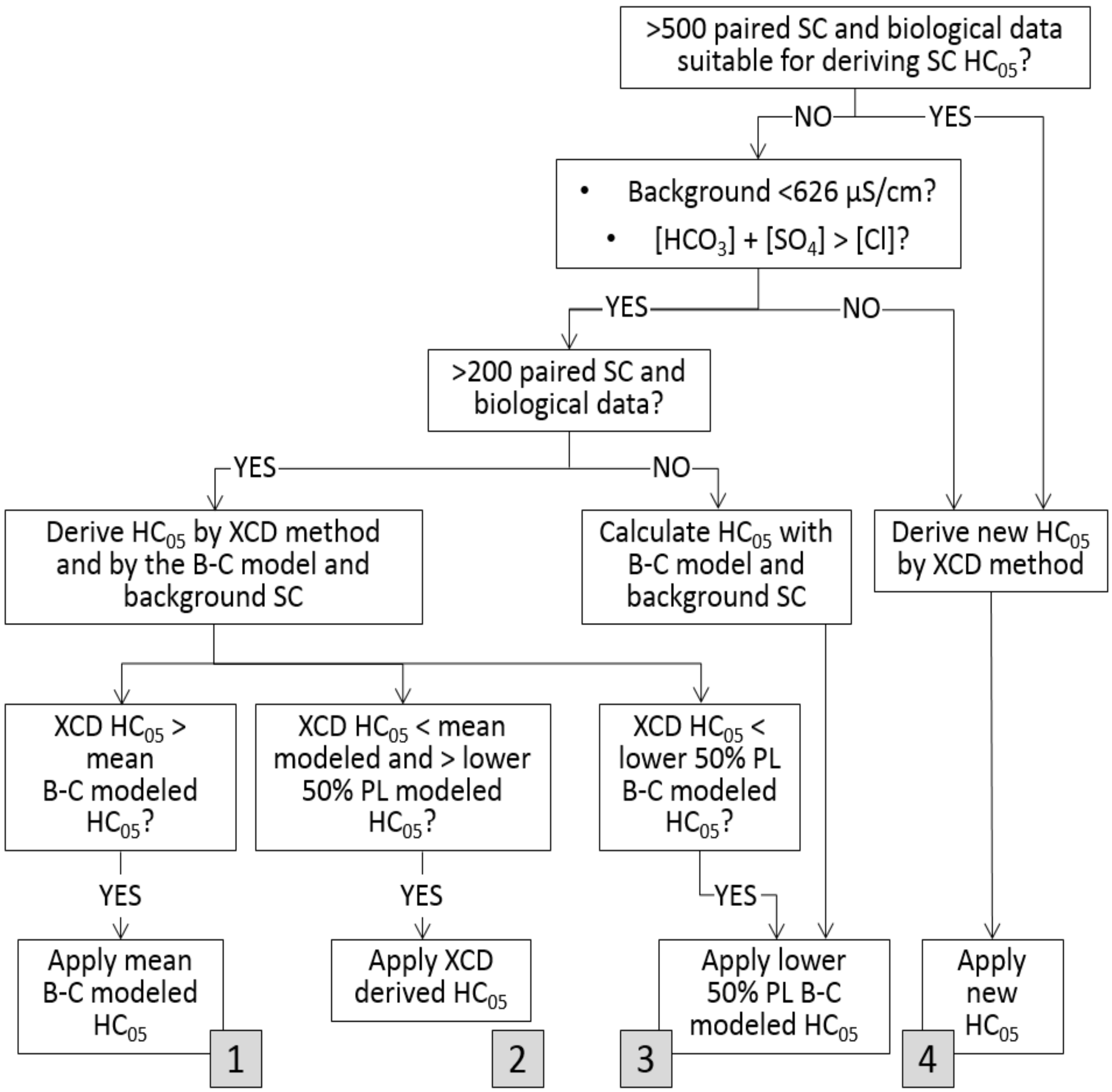

Figure D-7. A decision tree for calculating and applying 5th centile hazardous concentrations (HC_{05}). This flow chart describes the development of a HC_{05} for a new ecoregion, a new area within an ecoregion, or other defined geographic area using the extirpation concentration distribution (XCD) method or background-to-criterion (B-C) method. Numbered product paths are described in the body of the text.

Where the background is less than 626 µS/cm and the waters have a similar ion composition to those used to derive the model, the B-C method can be used (see Figure D-7).

Where there are >200 but <500 sites with paired biological and SC data, HC_{05} values are derived using the field-XCD method and compared to the mean and lower 50% PL of the the B-C model. These values are compared to select the appropriate HC_{05} as follows. (1) If the field-XCD HC_{05} (see eq 3-1) is greater than the mean B-C modeled HC_{05} (see eq D-1), then the mean B-C modeled HC_{05} is recommended (see eq D-1) as a conservative approach to account for uncertainty associated with a smaller data set. (2) If the field-XCD HC_{05} is between the mean B-C modeled HC_{05} and the lower 50% PL, then the field-XCD HC_{05} is recommended because the XCD from measured data from the region is more likely to represent the region than the more general B-C model. (3) If the field-XCD estimate is below the lower 50% PL, then the lower 50% PL is recommended as the HC_{05} (see eq D-2). Because the XCD is calculated from a smaller data set, it may be overly protective and is more uncertain than the modeled results which indicate that 75% of HC_{05} values from areas with a similar background SC are estimated to be greater than that value. Also, the lower 50% PL is also recommended when there are fewer than 200 paired biological samples because there is no XCD for comparison. The SC data and the B-C model is used to estimate the HC_{05}. (4) Where the background SC is greater than 626 µS/cm, the range of the model is exceeded, and it is recommended that data be collected to derive the HC_{05} using the XCD method (see Section 3.1).

The mean B-C model HC_{05} is calculated from eq D-1. The upper and lower PL for a predicted log10 HC_{05} value $\hat{y}$ can be calculated from the regression line using eq D-2 (Zaiontz, 2014) and log10 transformed SC values (x) as follows:

$$\hat{y} \pm t_{\alpha/2,n-2} S_y \sqrt{1 + \frac{1}{n} + \frac{(x^{\circ} - \bar{x})^2}{SS}} = PL \qquad \text{(D-2)}$$

Symbol	**Explanation**	**Example from the B-C model**
$\hat{y}$	Log10 of the mean predicted HC_{05}	Variable differs for each case
n	Number of samples	$n = 24$
α	Alpha error rate for prediction interval (desired confidence level)	50% prediction interval (α = 0.5)
t_{n-2}	Student's *t*-value at specified confidence level (alpha, α) and *n*-2 degrees of freedom	For 50% prediction interval (α = 0.5), $t_{(1-0.5)/2,24-2} = 0.686$

Symbol	Explanation	Example from the B-C model
S_y	Residual standard error of prediction (standard deviation)	S_y = 0.11
SS	Sum of square of *x* deviation from their mean, SS = $\sum_{i=1}^{n}(x_i - \bar{x})^2$	*SS* = 4.21
$\bar{x}$	Mean *x* values used in the model generation	$\bar{x}$ = 2.15
x°	A new log10 background *(x)* value for a new prediction interval	SC value differs for each case
PL	Upper and lower prediction limits of mean predicted $\hat{y}$	SC value differs for each case

The background values listed in Tables D-3, D-4 and D-5 for 62 of 85 Level III ecoregions were used to estimate the HC_{05} using the B-C model. Using those background values from the EPA-survey data with eqs 1 and 2 and estimated background from EPA-survey data, the HC_{05} values and PLs have been estimated for 62 of the 85 Level III ecoregions in the conterminous United States. Although the B-C Model is strongly log-linear within the sampled SC range, estimation of HC_{05} is recommended only for ecoregions with a background <626 µS/cm to avoid extrapolation beyond modeled data. Some regions may have different ionic matrices (e.g., sodium chloride-dominant) for which the derivation of a HC_{05} using this method has not been verified. Those ecoregions are identified in Table D-6. HC_{05} values estimated at the lower 50% PL and at the mean regression line are also provided in Tables D-3, D-4 and D-5. Because these are example HC_{05}s, the reported values have not been rounded. A CCC is the HC_{05} rounded to 2 significant digits.

The process described in Figure D-7 was used to select an HC_{05} value from among three different potential derivations: by the XCD method, or the mean or lower 50% PLs from the B-C model. Table D-3 provides lists of example HC_{05} values estimated with large data sets with well-defined background SC. Table D-4 lists ecoregions with background estimates based on modest survey data sets (*N* = 20 to 59 SC samples) that would benefit from additional sampling or comparison with state data sets to confirm the calculated background SC and the calculated HC_{05}. Table D-5 lists example HC_{05} values that support least disturbed conditions. These areas would also benefit from additional sampling or comparison with state data sets to confirm the calculated background SC and the calculated HC_{05}.

Table D-6 lists ecoregions that may not be served by the B-C model because the ionic pollutant is different, the natural background exceeds the range of the model, or the available survey data contained fewer than 20 samples.

The example modeled HC_{05} values for SC calculated with this B-C method should be considered provisional and states are encouraged to confirm that their ionic mixture is similar to those ecoregions used to develop the model. States are also encouraged to assess whether the background SC used to estimate a HC_{05} for an ecoregion is reliable and representative of an ecoregion. This is necessary because some background estimates are based on small sample sizes and because some ecoregions are large and may have different SC regimes in discrete areas of those ecoregions. If there are clear differences in background SC or ionic composition within an ecoregion, these areas should be differentiated from one another and independent calculations should be performed to yield appropriate and distinct HC_{05} values. Section 3.7.1.4 provides an approach for evaluating background SC and Appendix C provides an example of a weight-of-evidence analysis to help assess areas with similar or distinct background SC.

Table D-3. Provisional 5th centile hazardous concentration (HC_{05}) values using the background-to-criterion (B-C) method for Level III ecoregions estimated from large data sets. Values are from extirpation concentration distribution (XCD) method, B-C method predicted HC_{05}, and prediction limits for Level III ecoregions where the 25th centile specific conductivity [SC]) of sampled EPA survey sites is <626 µS/cm. Ecoregional stream SC 25th centiles are from EPA survey data sets. Figure D-4 was used to assign provisional values (gray). These values are well supported by large data sets for estimating background SC. Comparison of background SC and a geophysical model suggests that the background likely represents minimally affected condition.

Level III ecoregion number and name		State *N*	State 25th centile µS/cm	State XCD HC_{05} µS/cm	EPA-survey *N*	EPA-survey 25th centile µS/cm	B-C mean µS/cm	B-C lower PL µS/cm	B-C upper PL µS/cm	Mean base-flow modeled SC (µS/cm)
1	Coast Range	-	-	-	251	53	161	135	193	99
3	Willamette Valley	-	-	-	87	62	180	150	215	94
4	Cascades[a]	-	-	-	537[a]	33	118	98	142	66
7	Central California Valley	-	-	-	82	99	244	205	292	411
11	Blue Mountains	-	-	-	135	85	220	184	263	161
15	Northern Rockies	614	29	142	29	22	90	74	109	100
16	Idaho Batholith	1,040	42	185	29	36	126	105	151	81
17	Middle Rockies	510	136	264	89	78	209	175	250	229
19	Wasatch and Uinta Mountains[b]	773	271	299	32	31	114	95	137	285
21	Southern Rockies	-		-	164	83	218	182	260	218
23	Arizona/New Mexico Mountains[b]	374	191	249	38	105	254	213	304	385
35	South Central Plains[c]	-	-	-	60	51	157	131	188	117
40	Central Irregular Plains	-	-	-	60	301	507	424	606	324

Table D-3. Provisional 5th centile hazardous concentration (HC_{05}) values using the background to-criterion (B-C) method for Level III ecoregions estimated from large data sets. Values are from extirpation concentration distribution (XCD) method, B-C method predicted HC_{05}, and prediction limits for Level III Ecoregions where the 25th centile specific conductivity [SC]) of sampled EPA survey sites is <626 µS/cm. Ecoregional stream SC 25th centiles are from EPA survey data sets. Figure D-4 was used to assign provisional values (gray). These values are well supported by large data sets for estimating background SC. Comparison of background SC and a geophysical model suggests that the background likely represents minimally affected condition. (continued)

Level III ecoregion number and name		State *N*	State 25th centile µS/cm	XCD HC_{05} µS/cm	EPA-survey *N*	EPA-survey 25th centile µS/cm	B-C mean µS/cm	B-C lower PL µS/cm	B-C upper PL µS/cm	Mean base-flow modeled SC (µS/cm)
45	Piedmont	665	68	138	333	42	139	116	167	73
50	Northern Lakes and Forests	734	108	320	151	111	262	220	313	231
52	Driftless Area[b]	344	534	655	73	392	603	503	723	406
58	Northeastern Highlands	383	50	212	118	40	134	112	161	64
60	Northern Allegheny Plateau[b]	562	132	248	113	71	197	165	236	112
62	North Central Appalachians	-	-	-	131	33	118	98	142	60
65	Southeastern Plains (lower)[c,d]	457	38	131	241	26	101	83	122	88
65	Southeastern Plains (upper)[b,c,d]	359	103	257	241	25	101	83	122	88
66	Blue Ridge	322	22	69	245	16	74	61	90	59
67	Ridge and Valley	926	59	154	522	46	148	123	177	161
69	Central Appalachians	1,661	94	305	281	46	147	122	176	94

Table D-3. Provisional 5th centile hazardous concentration (HC_{05}) values using the background to-criterion (B-C) method for Level III ecoregions estimated from large data sets. Values are from extirpation concentration distribution (XCD) method, B-C method predicted HC_{05}, and prediction limits for Level III Ecoregions where the 25th centile specific conductivity [SC]) of sampled EPA survey sites is <626 µS/cm. Ecoregional stream SC 25th centiles are from EPA survey data sets. Figure D-4 was used to assign provisional values (gray). These values are well supported by large data sets for estimating background SC. Comparison of background SC and a geophysical model suggests that the background likely represents minimally affected condition. (continued)

Level III ecoregion number and name		State *N*	State 25th centile µS/cm	XCD HC_{05} µS/cm	EPA-survey *N*	EPA-survey 25th centile µS/cm	B-C mean µS/cm	B-C lower PL µS/cm	B-C upper PL µS/cm	Mean base-flow modeled SC (µS/cm)
70	Western Allegheny Plateau	2,075	169	338	109	153	325	272	388	180
77	North Cascades	-	-	-	73	27	105	87	126	65
78	Klamath Mountains/California High North Coast Range	-	-	-	74	83	218	182	260	150

[a]Data set is a combined EPA survey and U.S. Geological Survey (USGS) data set described in Section 7 of the main document.
[b]Difference of >50 µS/cm between state and EPA-survey background SC suggests need for further analysis.
[c]($[Na^+] + [K^+]$) and/or $[Cl^-]$ in mg/L is dominant, ≥50% of EPA-survey sample.
[d]Elongated ecoregion suggests need for further analysis.

Table D-4. Example 5th centile hazardous concentration (HC_{05}) values for Level III ecoregions using the background-to-criterion (B-C) method estimated from small data sets. These values are supported by moderately sized data sets (N = 23−60) for estimating background specific conductivity (SC). Therefore, these values are provided as a first approximation of HC_{05} values. Additional sampling or comparison with state data sets to confirm the calculated background SC and the calculated HC_{05} is recommended. Figure D-7 was used to assign example HC_{05} values (gray cells). Comparison of measured SC and a geophysical model suggests that the background represents minimally affected condition.

Level III ecoregion number and name		*N*	EPA survey 25th centile µS/cm	B-C Mean µS/cm	B-C lower PL µS/cm	B-C upper PL µS/cm	Mean base flow modeled SC (µS/cm)
6	Central California Foothills and Coastal Mountains[a]	27	244	441	369	527	480
8	Southern California Mountains	45	262	462	386	552	511
9	Eastern Cascades Slopes and Foothills	45	50	156	130	187	148
10	Columbia Plateau	26	128	288	242	344	264
13	Central Basin and Range	51	152	323	270	385	367
18	Wyoming Basin	37	397	607	507	728	443
20	Colorado Plateaus	33	436	646	538	774	495
25	High Plains	40	359	569	475	681	452
26	Southwestern Tablelands	25	495	703	585	844	544
29	Cross Timbers	23	288	492	411	588	384
36	Ouachita Mountains[a]	50	22	91	75	110	102
37	Arkansas Valley[a]	47	32	116	97	140	163
38	Boston Mountains	26	23	94	77	113	187
40	Central Irregular Plains	60	301	507	424	606	324
42	Northwestern Glaciated Plains	46	342	551	460	659	421
44	Nebraska Sand Hills	34	161	335	281	400	263
63	Middle Atlantic Coastal Plain	59	93	235	196	280	96
68	Southwestern Appalachians	55	40	133	111	160	104
74	Mississippi Valley Loess Plains[a]	26	69	192	161	230	110

Table D-4. Example 5th centile hazardous concentration (HC_{05}) values for Level III ecoregions using the background-to-criterion (B-C) method estimated from small data sets. These values are supported by moderately sized data sets (*N* = 23−54) for estimating background specific conductivity (SC). Therefore, these values are provided as a first approximation of HC_{05} values. Additional sampling or comparison with state data sets to confirm the calculated background SC and the calculated HC_{05} is recommended. Figure D-7 was used to assign example HC_{05} values (gray cells). Comparison of measured SC and a geophysical model suggests that the background represents minimally affected condition. (continued)

Level III ecoregion number and name		*N*	EPA survey 25th centile µS/cm	B-C Mean µS/cm	B-C lower PL µS/cm	B-C upper PL µS/cm	Mean base flow modeled SC (µS/cm)
75	Southern Coastal Plain[a]	50	52	160	133	191	109
80	Northern Basin and Range	47	94	236	198	282	269
82	Acadian Plains and Hills	23	53	162	135	194	89
84	Atlantic Coastal Pine Barrens[a]	37	62	180	150	215	68

[a] ($[Na^+] + [K^+]$) and/or $[Cl^-]$ in mg/L is dominant, ≥50% of EPA-survey sample.

Table D-5. Example values where 25th centile specific conductivity (SC) likely represents least disturbed conditions. Example measured hazardous concentration (HC_{05}) values using the extirpation concentration distribution (XCD) method and calculated values using background-to-criterion (B-C) method. Range in parentheses where sample size was <20 are included in this table because a larger paired data set was available from another source. Data from Figure D-3 was used to assess predicted mean base-flow SC. Figure D-7 was used to assign example HC_{05} values (gray). The measured 25th centile background SC is greater than the mean geophysical model suggesting that the background represents least disturbed conditions.

Level III ecoregion number and name		*State N*	State 25th centile µS/cm	State XCD HC_{05} µS/cm	EPA-Survey *N*	EPA-Survey 25th centile µS/cm	B-C mean µS/cm	B-C lower PL µS/cm	B-C upper PL µS/cm	Mean base-flow modeled SC (µS/cm)
27	Central Great Plains	-	-	-	133	469	678	565	814	452
39	Ozark Highlands	-	-	-	54	362	571	477	684	278
43	Northwestern Great Plains[a,b]	-	-	-	399[b]	542	745	620	896	489
47	Western Corn Belt Plains[c]	473	587	934	178	480	688	573	826	349
51	North Central Hardwood Forests[c]	583	325	494	35	149	319	267	381	301
54	Central Corn Belt Plains[c]	465	626	813	14	465 (465−1,100)	674	562	809	267
55	Eastern Corn Belt Plains	-	-	-	25	612	808	671	972	247
59	Northeastern Coastal Zone[a,c]	277	350	706	41	105	254	213	303	72
61	Erie Drift Plain	-	-	-	33	230	424	355	507	166
64	Northern Piedmont	539	150	257	92	114	268	224	320	101
71	Interior Plateau	336	296	479	56	299	504	421	603	236
72	Interior River Valleys and Hills[c]	460	460	1,108	51	384	594	496	712	264

Table D-5. Example values where 25th centile specific conductivity (SC) likely represents least disturbed conditions. Example measured hazardous concentration (HC_{05}) values using the extirpation concentration distribution (XCD) method and calculated values using background-to-criterion (B-C) method**.** Range in parentheses where sample size was <20 are included in this table because a larger paired data set was available from another source. Data from Figure D-3 was used to assess predicted mean base-flow SC. Figure D-7 was used to assign example HC_{05} values (gray). The measured 25th centile background SC is greater than the mean geophysical model suggesting that the background represents least disturbed conditions. (continued)

Level III ecoregion number and name		***State N***	**State 25th centile µS/cm**	**State XCD HC_{05} µS/cm**	**EPA-Survey *N***	**EPA-Survey 25th centile µS/cm**	**B-C mean µS/cm**	**B-C lower PL µS/cm**	**B-C upper PL µS/cm**	**Mean base-flow modeled SC (µS/cm)**
73	Mississippi Alluvial Plain	-	-	-	27	132	294	246	351	121
83	Eastern Great Lakes Lowlands[c]	591	272	525	17	104 (104−1,890)	251	211	300	181

[a]($[Na^+] + [K^+]$) and/or $[Cl^-]$ in mg/L is dominant, ≥50% of EPA-survey sample.
[b]Data set is a combined EPA survey and USGS data set described in Section 6 of the main document.
[c]Difference of >50 µS/cm between state and EPA-survey background SC suggests need for further analysis.

Table D-6. Level III ecoregions with uncalculated 5th centile hazardous concentrations (HC_{05}). The 25th centile background was outside the model's range (background >626 µS/cm) (exceeds) or the sample size to estimate the 25th centile background was <20 samples (Sm). Ecoregion 76, the Everglades, is a unique system and is influenced by salt water intrusion. Additional data would be needed to calculate a HC_{05} for these ecoregions.

Level III ecoregion number and name		Qualifier	*N*	EPA-survey range µS/cm	Mean base flow modeled SC (µS/cm)
2	Puget Lowland	Sm	11	(50−228)	71
5	Sierra Nevada	Sm	18	(19.1−207)	144
12	Snake River Plain	Sm	4	(1,365−582)	300
14	Mojave Basin and Range	Exceeds	2	(2,860)	503
22	Arizona/New Mexico Plateau	Sm	10	(76.9−533)	486
24	Chihuahuan Deserts	Exceeds	1	(3,500)	504
28	Flint Hills	Sm	18	(350−800)	472
30	Edwards Plateau	Sm	6	(372−693)	489
31	Southern Texas Plains	No data	-	-	462
32	Texas Blackland Prairies	Sm	5	(143−849)	321
33	East Central Texas Plains	Sm	6	(188−1,560)	254
34	Western Gulf Coastal Plain	Sm	16	(71−2,046)	203
41	Canadian Rockies	Sm	7	(117−364)	138
46	Northern Glaciated Plains	Exceeds	21	(451−2,890)	329
48	Lake Agassiz Plain Range	Sm	13	(618−2,630)	341
49	Northern Minnesota Wetlands	Sm	8	(77.3−607)	256
53	Southeastern Wisconsin Till Plains Range	Sm	9	(227−1,070)	388
56	Southern Michigan/Northern Indiana Drift Plains	Sm	19	(295−1,140)	246
57	Huron/Erie Lake Plains	Sm	8	(336−1,980)	307
76	Southern Florida Coastal Plain	Everglades	603	(1.2−4,540)	183
79	Madrean Archipelago	No data	-	-	401
81	Sonoran Basin and Range	Sm	5	(279−12,300)	495
85	Southern California/Northern Baja Coast	Sm	12	(368−4,100)	566

D.7. REFERENCES

D.7.1. Data Sources for Predictor Variables for Mean Base-Flow Specific Conductivity

Variable	Name	Source
BDH_AVE	Mean soil bulk density	https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/geo/?cid=nrcs142p2_053627
EVI_JAS	Enhanced vegetation index	http://lpdaac.usgs.gov
KFAC_AVE	Mean soil erodibility	https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/geo/?cid=nrcs142p2_053627
LST32_AVE	Mean day last freeze	Daly et al., 1994
PCT_EvGr	Percentage evergreen	http://www.mrlc.gov/
PermH_AVE	Mean soil permeability	https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/geo/?cid=nrcs142p2_053627
MinP_AVE	Mean annual minimum monthly temperature	Daly et al., 1994
MeanP_AVE	Mean precipitation	Daly et al., 1994
Tmax_AVE	Mean annual maximum monthly temperature	Daly et al., 1994
XWD_AVE	Mean number of wet days	Daly et al., 1994
AtmCa_Ave	Mean atmospheric calcium deposition	Olson and Hawkins, 2012
CaO_Mean	Mean whole rock CaO	Olson and Hawkins, 2012
S_Mean	Mean whole rock S	Olson and Hawkins, 2012
UCS_Mean	Mean unconfined compressive strength	Olson and Hawkins, 2012

D.7.2. Literature Cited

Arlot, S; Cleisse, A. (2010) A survey of cross-validation procedures for model selection. Stat Surv 4:40−79. Available online at http://projecteuclid.org/download/pdfview_1/euclid.ssu/1268143839.

Brinck, EL; Drever, JI; Frost, CD. (2008) The geochemical evolution of water coproduced with coalbed natural gas in the Powder River Basin, Wyoming. Environ Geosci 15(4):153−171.

Clark, ML; Miller, KA; Brooks, MH. (2001) U.S. Geological Survey monitoring of Powder River Basin stream water quantity and quality. Cheyenne, WY: U.S. Geological Survey, U.S. Department of the Interior, Cheyenne, WY. Available online at http://pubs.usgs.gov/wri/wri014279/.

Dahm, KG; Guerra, KL; Xu, P; Drewes, JE. (2011) Composite geochemical database for coalbed methane produced water quality in the Rocky Mountain region. Environ Sci Technol 45(18):7655−7663.

Daly, C; Neilson, RP; Phillips, DL. (1994) A statistical topographic model for mapping climatological precipitation over mountainous terrain. J Appl Meteorol 33:140−158.

Entrekin, S; Evans-White, M; Johnson, B; Hagenbuch, E. (2011) Rapid expansion of natural gas development poses a threat to surface waters. Front Ecol Environ 9:503−511.

Environment Canada and Health Canada. (2001) Priority substances list assessment report: road salts. Canadian Environmental Protection Act, 1999. Available online at http://ww1.prweb.com/prfiles/2008/02/07/370423/EnvironmentCanadareport.pdf.

ESRI (Environmental Systems Resource Institute) (2012) ArcMap 10.1. Relands, CA: ESRI. Available online at www.esri.com.

Evans, M; Frick, C. (2001) The effects of road salts on aquatic ecosystems. NWRI Contribution Series No. 02-308. Saskatoon, Saskatchewan: National Water Research Institute (NWRI).

Forman, RTT; Alexander, LE. (1998) Roads and their major ecological effects. Ann Rev Ecol Syst 29:207−31.

Gregory, KB; Vidic, RD; Dzombak, DA. (2011) Water management challenges associated with the production of shale gas by hydraulic fracturing. Elements 7(3):181−186.

Haluszczak, LO; Rose, AW; Kump, LR. (2013) Geochemical evaluation of flowback brine from Marcellus gas wells in Pennsylvania, USA. Appl Geochem 28:55−61.

Hill, RA; Weber, MH; Leibowitz, SG; Olsen, AR; Thornbrugh, DJ. (2016) The Stream-Catchment (StreamCat) Dataset: A database of watershed metrics for the conterminous United States. JAWRA 52:120-128.

Jackson, RE; Reddy, KJ. (2007) Geochemistry of coalbed natural gas (CBNG) produced water in Powder River Basin, Wyoming: salinity and sodicity. Water Air Soil Poll 184:49−61.

James, G; Witten, D; Hastie, T; Tibshirani, R. (2013) An introduction to statistical learning with applications in R. New York, NY: Springer Science and Business Media. Available on line at http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Fourth%20Printing.pdf.

Kaushal, SS; Groffman, PM; Likens, GE; Belt, KT; Stack, WP; Kelly, VR; Band, LE; Fisher, GT. (2005) Increased salinization of fresh water in the northeastern United States. Proc Natl Acad Sci USA 102(38):13517−13520.

Kelly, VR; Lovett, GM; Weathers, KC; Findlay, SEG; Strayer, DL; Burns, DI; Likens, GE. (2008) Long term sodium chloride retention in a rural watershed: legacy effects of road salt on streamwater concentration. Environ Sci Technol 42(2):410−415.

Kunz, JL; Conley, JM; Buchwalter, DB; Norber-King, TJ; Kemble, NE; Wang, N; Ingersoll, CG. (2013) Use of reconstituted waters to evaluate effects of elevated major ions associated with mountaintop coal mining on freshwater invertebrates. Environ Toxicol Chem 32(12):2826−2835.

Mount, DR; Erickson, RJ; Highland, TL; Hockett, JR; Hoff, DJ; Jenson, CT; Norberg-King, TJ; Peterson, KN; M. Polaske,ZM; Wisniewskiz, S. (2016) The acute toxicity of major ion salts to Ceriodaphnia dubia: I. Influence of background water chemistry. Environ Toxicol Chem 35(12):3039-3057.

NRC (National Research Council). (2010) Management and effects of coalbed methane produced water in the western United States. Committee on Management and Effects of Coalbed Methane Development and Produced Water in the Western United States. Washington, D.C.: National Academies Press.

Olson, JR; Hawkins, CP. (2012) Predicting natural base-flow stream water chemistry in the western United States. Water Resour Res 48:W02504.

Omernik, JM. (1987) Ecoregions of the conterminous United States. Ann Assoc Am Geograph 77(1):118−125.

Omernik, JM. (1995) Ecoregions: a spatial framework for environmental management. In: Davis, WS; Simon, TP; (eds). Biological assessment and criteria: tools for water resource planning and decision making. Pp 49−62 Boca Raton, FL: Lewis Publishers.

Soucek, DJ; Dickinson, A. (2015) Full-life chronic toxicity of sodium salts to the mayfly Neocloeon triangulifer in tests with laboratory cultured food. Environ Toxicol Chem 34(9):2126-37.

Stoddard, JL; Larsen, DP; Hawkins, CP; Johnson, RK; Norris, RH. (2006) Setting expectations for the ecological condition of streams: The concept of reference condition. Ecol Appl 16(4):1267−1276.

Veil, JA; Puder, MG; Elcock, D; Redweik, RJ, Jr. (2004) A white paper describing produced water from production of crude oil, natural gas, and coal bed methane. Argonne, IL: U.S. Department of Energy, Argonne National Laboratory. Available on-line at: http://citeseerx.ist.psu.edu/viewdoc/download?rep=rep1&type=pdf&doi=10.1.1.211.8898.

Zaiontz, C. (2014) Real statistics using Excel. Available online at http://www.real-statistics.com/regression/confidence-and-prediction-intervals/.