Environmental Toxicology and Chemistry—Volume 42, Number 2—pp. 393–413, 2023
Received: 27 May 2022 | Revised: 7 July 2022 | Accepted: 14 November 2022

Environmental Toxicology

Comparison of Multiple Linear Regression and Biotic Ligand Models for Predicting Acute and Chronic Zinc Toxicity to Freshwater Organisms

David K. DeForest,[a,*] Adam C. Ryan,[b] Lucinda M. Tear,[a] and Kevin V. Brix[c,d]

[a]Windward Environmental, Seattle, Washington, USA
[b]International Zinc Association, Durham, North Carolina, USA
[c]EcoTox, Miami, Florida, USA
[d]Rosenstiel School of Marine and Atmospheric Science, University of Miami, Miami, Florida, USA

Abstract: Multiple linear regression (MLR) models for predicting zinc (Zn) toxicity to freshwater organisms were developed based on three toxicity-modifying factors: dissolved organic carbon (DOC), hardness, and pH. Species-specific, stepwise MLR models were developed to predict acute Zn toxicity to four invertebrates and two fish, and chronic toxicity to three invertebrates, a fish, and a green alga. Stepwise regression analyses found that hardness had the most consistent influence on Zn toxicity among species, whereas DOC and pH had a variable influence. Pooled acute and chronic MLR models were also developed, and a *k*-fold cross-validation was used to evaluate the fit and predictive ability of the pooled MLR models. The pooled MLR models and an updated Zn biotic ligand model (BLM) performed similarly based on (1) R^2, (2) the percentage of effect concentration (EC*x*) predictions within a factor of 2.0 of observed EC*x*, and (3) residuals of observed/predicted EC*x* versus observed EC*x*, DOC, hardness, and pH. Although fit of the pooled models to species-specific toxicity data differed among species, species-specific differences were consistent between the BLM and MLR models. Consistency in the performance of the two models across species indicates that additional terms, beyond DOC, hardness, and pH, included in the BLM do not help explain the differences among species. The pooled acute and chronic MLR models and BLM both performed better than the US Environmental Protection Agency's existing hardness-based model. We therefore conclude that both MLR models and the BLM provide an improvement over the existing hardness-only models and that either could be used for deriving ambient water quality criteria. *Environ Toxicol Chem* 2023;42:393–413. © 2022 SETAC

Keywords: Ambient water quality criteria; bioavailability; biotic ligand model; multiple linear regression; zinc

INTRODUCTION

The US Environmental Protection Agency's (USEPA's) current acute and chronic ambient water quality criteria (AWQC) for zinc (Zn) in freshwater were developed in 1987 (USEPA, 1987), with an update to include new toxicity data in 1995 (USEPA, 1996). The freshwater Zn criteria are adjusted for water hardness because it reduces Zn toxicity. This is a result of competition between calcium and Zn for uptake by aquatic organisms (Hogstrand et al., 1998; Spry & Wood, 1989). Because limited chronic Zn toxicity data were available at the time, the USEPA's chronic Zn criteria were derived using an acute-to-chronic ratio.

Over the last 25 years, Zn toxicity data for freshwater organisms have been expanded in two significant ways. First, the number and diversity of species for which Zn toxicity has been assessed in chronic exposures have increased substantially such that chronic criteria, following USEPA methods (USEPA, 1985), may now be developed directly from empirical chronic toxicity data without the use of an acute-to-chronic ratio. Second, both acute and chronic toxicity testing with Zn have been conducted over varying water chemistry conditions that influence Zn bioavailability. This testing supported the development of biotic ligand models (BLMs) for predicting Zn toxicity to cladocerans (*Daphnia magna*, *D. pulex*), fathead minnow (*Pimephales promelas*), rainbow trout (*Oncorhynchus mykiss*), and a green alga (*Raphidocelis subcapitata*, formerly *Pseudokirchneriella subcapitata*; Clifford & McGeer, 2009;

This article includes online-only Supporting Information.
* Address correspondence to davidd@windwardenv.com
Published online 18 November 2022 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/etc.5529

De Schamphelaere & Janssen, 2004; De Schamphelaere et al., 2003, 2005; Heijerick et al., 2002, 2005; Santore et al., 2002). A "unified" Zn BLM was subsequently developed based on a synthesis of existing Zn BLMs and BLM-based Zn criteria were developed following USEPA guidelines, using updated acute and chronic datasets (DeForest & Van Genderen, 2012).

More recently, empirical multiple linear regression (MLR) models have been developed for predicting metal toxicity as a function of multiple water chemistry variables (e.g., dissolved organic carbon [DOC], hardness, pH; Brix et al., 2017, 2021; Croteau et al., 2021; DeForest et al., 2018, 2020; Peters et al., 2021; Stauber et al., 2021), and the USEPA has applied this modeling framework in an update to the AWQC for aluminum (Al; USEPA, 2018). The relative performance of MLR models and BLMs for copper (Cu; Brix et al., 2021) and nickel (Ni; Croteau et al., 2021) have recently been compared following the methods presented in Garman et al. (2020) and Brix et al. (2020). The objectives of the present study were to (1) develop acute and chronic MLR models for Zn, (2) optimize the Zn BLM based on the same datasets used to develop the MLR models, and (3) compare the performance of the MLR models to the Zn BLM.

METHODS

Toxicity database

A comprehensive Zn toxicity database was developed. The types of toxicity data compiled were not specific to guidance for any particular regulatory jurisdiction, and therefore the database was not constrained to include only specific endpoints and exposure durations. However, for the purposes of comparing MLR models and the BLM, an emphasis was placed on evaluating toxicity data that generally met USEPA (1985) guidelines for AWQC development.

Most acute toxicity data were from tests 48 to 96 h in duration with mortality (or mortality and immobilization) as the endpoint, although 24-h tests were considered acceptable for rotifers due to their very short life cycle (ASTM International, 2004). Acute toxicity data expressed as median lethal concentrations (LC50s) and median effect concentrations (EC50s) were compiled.

With the exceptions noted below, relevant chronic toxicity data consisted of life cycle exposures for invertebrates and early life stage (ELS), partial life cycle, or full life cycle exposures for fish. Exceptions to full life cycle tests for invertebrates included ELS exposures with snails and mussels, and exceptions for fish included shorter-term exposures with salmonids (such as 30-day juvenile survival and growth tests based on evidence that these tests had a sensitivity similar to longer-term ELS tests). See Supporting Information, Data S1, for the rationale on the toxicity tests and endpoints that were retained despite not strictly meeting (USEPA, 1985) guidelines for AWQC development.

The preferred statistical endpoints from chronic tests were EC10s, EC20s, and EC50s. When these values were not reported, they were derived independently using the Toxicity Relationship Analysis Program (TRAP, Ver. 1.30a; USEPA, 2015) if raw data were provided. Where it was necessary to calculate at least one of the preferred statistical endpoints from raw data, we also preferentially used the re-calculated effect concentration (EC*x*) from the same concentration-response evaluation (i.e., we replaced reported EC*x* with recalculated values, as necessary). Historically, the USEPA has used EC20s as chronic toxicity thresholds for deriving AWQC (USEPA, 2013, 2016), but chronic EC10s were also compiled because this is the preferred effect level for other jurisdictions (Canadian Council of Ministers for the Environment, 2007; European Chemicals Agency, 2008; Warne et al., 2018). We also compiled chronic EC50s in case evaluation of these values facilitated interpretation of models developed from lower effect concentrations.

Although algae and aquatic plants are not directly incorporated into calculation of USEPA's AWQC, assessing the protectiveness of AWQC based on aquatic animals (i.e., invertebrates, fish, amphibians) relative to the sensitivity of algae and aquatic plants is a step in the AWQC development process (USEPA, 1985). Consequently, and consistent with other jurisdictions that directly consider plant data in setting water quality criteria, we also compiled Zn toxicity data for algae and aquatic plants.

For MLR model development, toxicity tests were compiled that had reported pH and hardness (or calcium [Ca] and magnesium [Mg]), and for which DOC was either reported or for which it could reasonably be estimated based on the source water. Estimations of DOC from source water were conducted according to Appendix C in USEPA (2007) or, in some cases, using professional judgment. For example, in chronic tests with daphnids, DOC will increase with addition of food. If the DOC concentration was reported only in the dilution water source prior to testing, or if DOC was estimated (e.g., for tests conducted in synthetic water), a DOC concentration of 2 mg/L was assumed in chronic daphnid tests (Besser et al., 2021).

Finally, in developing the Zn toxicity database, tests that used a pH buffer were noted and potential implications on the MLR models and BLM were assessed by evaluating whether there were differences in pH-dependent effects in tests with and without buffers. In addition, test organism life stages were documented and potential life stage effects on sensitivity were evaluated once initial models allowed for normalization to common water chemistry conditions (i.e., so potential life stage effects could be differentiated from variations in bioavailability conditions).

MLR models

Based on previous Zn toxicity studies, and in particular studies involving development of Zn BLMs, DOC, hardness, and pH were identified as the toxicity modifying factors (TMFs) having the greatest potential influence on Zn bioavailability. Following the recommendation in Brix et al. (2017), acute and chronic Zn MLR models required toxicity tests with a minimum DOC range of 5 mg/L, a minimum hardness range of 100 mg/L, and a minimum pH range of 1.5 pH units. For acute toxicity, *Ceriodaphnia dubia*, *D. magna*, *D. pulex*, *O. mykiss*, *P. promelas*, and the snail *Pomacea paludosa* met these criteria.

For chronic toxicity, *C. dubia*, *D. magna*, the snail *Lymnaea stagnalis*, *O. mykiss*, and *R. subcapitata* had sufficient data to meet these criteria. Acute models were developed based on EC50s and chronic models were developed based on EC10s, EC20s, and EC50s.

Stepwise MLR analyses were conducted using an initial model that included ln(DOC), ln(hardness [Hard]), and pH as independent variables and ln(EC*x*) as the dependent variable. The first iteration of the model analysis assumed that ln(DOC), ln(Hard), and pH do not interact in their influence on Zn bioavailability and toxicity. The general form of the MLR model without TMF interaction terms is:

$$\ln(ECX) = b_0 + b_1 \times \ln(DOC) + b_2 \times \ln(Hard) + b_3 \times pH \quad (1)$$

where ECX is EC10, EC20, or EC50 (μg Zn/L), b_0 is the intercept, b_1, b_2, and b_3 are the slopes for ln(DOC), ln(Hard), and pH, respectively, and DOC and hardness are in units of mg/L.

A second iteration of the model analysis was conducted to evaluate whether the main independent variables interact with each other to modify the bioavailability and toxicity of Zn. In this analysis, the general form of the model includes the main independent variables as well as all two-way interactions:

$$\begin{aligned}\ln(ECX) = {} & b_0 + b_1 \times \ln(DOC) + b_2 \times \ln(Hard) + b_3 \times pH + b_4 \\ & \times \ln(DOC) \times pH + b_5 \times \ln(DOC) \times \ln(Hard) \\ & + b_6 \times ln(Hard) \times pH\end{aligned} \quad (2)$$

where ECX, b_0, b_1, b_2, and b_3 are defined as above, and b_4, b_5, and b_6 are the slopes for ln(DOC) × pH, ln(DOC) × ln(Hard), and ln(Hard) × pH, respectively.

Acute and chronic MLR models were developed for individual species and also by pooling all species within each exposure type (i.e., a pooled acute model and a pooled chronic model). Pooled chronic models were evaluated with and without inclusion of the alga *R. subcapitata* because TMFs appear to influence Zn bioavailability and toxicity to algae differently than for invertebrates and fish (De Schamphelaere et al., 2005). These pooled MLR models were evaluated using the approach described for Cu in Brix et al. (2017, 2021). In the pooled models, the combined data were used to fit common slopes for parameters found to have an important influence on Zn bioavailability and toxicity (based on the Akaike information criterion [AIC] and the Bayesian information criterion [BIC]), with species-specific intercepts to account for differences in species' sensitivities.

In species-specific modeling, AIC and BIC were used to determine the combination of terms that resulted in the most parsimonious models for predicting Zn EC*x* for each species. Stepwise regression analyses were conducted in R Ver. 4.0.4 (R Core Team, 2021) using the stepAIC() function from the MASS library (for the BIC model, stepAIC(model, direction = "both," $k = 1n(n)$), where n = sample size; Venables & Ripley, 2002). Variance inflation factors (VIFs) were calculated in R using the vif() function in the usdm library (Naimi et al., 2014).

To evaluate model performance, adjusted R^2 was calculated to assess the proportion of the variability in Zn toxicity values that could be explained by the model, adjusting for the number of terms in the model and the number of discrete observations. Predicted R^2 was calculated to assess the ability of the model to predict new data, substituting the Predictive Residual Sums of Squares (PRESS) for the Total Residual Sums of Squares in the R^2 calculation. The predicted R^2 will always be at least slightly lower than the adjusted R^2; large decreases in the predicted R^2 relative to the adjusted R^2 indicate that the model may be overparameterized or strongly influenced by one or more tests in the dataset. Although no a priori criteria exist to determine how large a difference is meaningful, we generally consider a predicted $R^2 > 0.10$ lower than the adjusted R^2 to merit further evaluation and a reduction >0.20 to indicate significant problems with the model.

MLR model validation

Validation of the MLR models in the present study was conducted differently from previous analyses we have conducted and other studies in the literature. For example, for metals with large datasets, such as Cu and Ni, validation datasets were identified a priori and used to validate the models but were not then subsequently incorporated back into the dataset to develop final models based on all data (Brix et al., 2021; Croteau et al., 2021). For metals with smaller datasets, such as Al, it was not possible to set aside data for model validation without compromising model development (DeForest et al., 2020).

In the present study, we used a *k*-fold cross-validation method that allows for both model validation and inclusion of all data in the final model (Hastie et al., 2009). Given the size of the pooled model datasets, we set $k = 5$ and conducted five different cross-validation runs, comparing the results of each run based on model errors (or fit) and model coefficients. To accomplish this, the dataset was randomly divided into five subsets ("folds"), each containing approximately 20% of the data (if a study included multiple tests, individual tests were randomly divided into the folds). Because data were randomly divided among folds, data for a given species were also randomly divided among folds. In each cross-validation run, a "training model" was fit to four of the five subsets (80% of the full dataset) and the training model was used to predict the results of the data in the unused "test" fold. In each run, a different fold (20%) was used as the test fold so that each of the original five folds eventually served as a test subset.

The acute and chronic full pooled invertebrate and fish species models and the chronic full alga (*R. subcapitata*) model (i.e., those including DOC, hardness, and pH with no interaction terms) were used as the basic training models in each cross-validation run. Model coefficients and fit parameters (predicted R^2 and root mean square error [RMSE]; Equation 3) were calculated for each of the five training models along with the RMSE of each test subset's fit to the training model. The mean, standard deviation, and coefficient of variation (CV) of each TMF coefficient and predicted R^2 were calculated across

the five training models to provide insight about how model coefficients and fit varied based on the data used.

$$\text{RMSE} = \sqrt{\frac{SS_{residuals}}{n}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \hat{x})^2}{n}} \quad (3)$$

where $SS_{residuals}$ is the sum of squared residuals, x_i = ln(observed ECx), and $\hat{x}$ = predicted ln(ECx).

In addition, because the ultimate objective in evaluating MLR models was to support development of bioavailability-based AWQC, we also developed MLR-based fifth percentile hazardous concentrations (HC5s) using the training models from each of the five folds in the cross-validation. This provided a measure of how Zn HC5s could vary across subsets of the toxicity data used to develop the models. Details on the methods for deriving HC5s are described in the ambient water quality criteria section because our study also included a comparison of HC5s based on the final recommended MLR models and updated Zn BLM.

Individual TMF regression analysis

Although the MLR analysis required minimum TMF ranges for each species modeled, previous MLR analyses on datasets for Cu and Ni have demonstrated data can be unevenly distributed across the range of a given TMF and may also be weighted more heavily toward tests in which one TMF was varied more frequently than others (Brix et al., 2021; Croteau et al., 2021). For example, for many species, more tests were conducted over a wide range of hardness conditions and fewer tests were conducted over wide ranges of DOC or pH. For DOC, it is not unusual for the majority of tests for a species to be conducted at a low DOC concentration (e.g., <2 mg/L) with only a small number of tests conducted at moderate to high DOC concentrations. Similarly, for pH, a higher proportion of tests tend to be conducted between a pH of 7 and 8 and a lower proportion in more acidic and basic waters. To evaluate whether any TMF slopes in the MLR models may be biased due to underrepresentation or uneven distribution in the dataset, univariate regression analyses were conducted for tests in which a single TMF was varied and the other two were held relatively constant.

The univariate regression analyses for individual TMFs were conducted by applying the same approach used by the USEPA to develop pooled hardness slopes to support hardness-based AWQC (USEPA, 1985). For each species used to develop the pooled acute and chronic MLR models for invertebrates and fish, subsets of tests were identified in which a single TMF was varied. Using those subsets of data, an analysis of covariance (Equation 4) was conducted to test for significant differences in the slope of the TMF among species (lm() and anova() functions in the stats package R Core Team (2021). The results of this analysis were used to support interpretation of the MLR models. If species-specific slopes did not differ (p value of sp × TMF interaction term >0.05), then a pooled slope term and species-specific intercepts were calculated (Equation 5):

$$\ln(\text{ECX}) = b_0 + b_1 \times \text{TMF} + b_2 \times \text{Species} + b_3 \times \text{TMF} \times \text{Species} \quad (4)$$

$$\ln(\text{ECX}) = b_0 + b_1 \times \text{TMF} + b_2 \times \text{Species} \quad (5)$$

where b_0 is the mean intercept, b_1 is the TMF slope, b_2 is the difference between mean intercept and intercept for species i, b_3 is the coefficient of interaction between TMF and species i (i.e., coding for species-specific slopes), TMF = ln(DOC), ln(Hard), or pH, and Species represents the coding variables for species i (based on matrix of contrasts).

Biotic ligand model

Parameter Estimation Software (PEST; Doherty, 2018) was used to update the BLM parameters for all subsets of data for which MLR models were developed. These included pooled models and species-specific models for both acute (EC50) and chronic (EC10, EC20, and EC50) subsets. The PEST performs parameter optimization by an iterative, least squares procedure, and provides a mechanism for objectively developing a set of BLMs that are parallel to the set of MLR models.

The PEST optimization procedure was conducted with log10-tranformed dissolved Zn effect concentrations (ECx). Only the biotic ligand parameters were optimized using PEST. The thermodynamic database specifying the bulk solution chemical reactions was not modified. To specify reasonable ranges of best-fitting biotic ligand parameters, we used Table 2 in DeForest & Van Genderen (2012) to guide definition of the ranges over which PEST would search for optimum BLM parameters. The ranges for each log K were tabulated and then expanded by one log unit at the lower and upper ends of the range. For example, the range of biotic ligand-Ca log Ks reported for BLMs in DeForest & Van Genderen (2012) was 3.2–4.9, therefore the range we used for PEST optimizations (for all fish and invertebrates) was 2.2–5.9 (Supporting Information, Table S2-1 in Data S2). For algae, we expanded the range further (Supporting Information, Table S2-1) because BLMs for algae were not addressed in DeForest & Van Genderen (2012). Previous work has demonstrated that generalized bioavailability models (gBAMs) can perform well for algae (De Schamphelaere et al., 2005; Van Regenmortel et al., 2017), but we wanted to evaluate if the parameters of the DeForest & Van Genderen (2012) "unified" BLM could be updated to adequately predict ECx for algae. Recognizing that application of the DeForest & Van Genderen (2012) "unified" BLM to algae data tended to overpredict ECx at high pH and underpredict ECx at low pH, we increased the starting value of the log K for biotic ligand-ZnOH in the PEST optimizations for algae.

For all PEST optimizations, two of the biotic ligand parameters (biotic ligand-Zn and biotic ligand-H) were used to anchor the optimization procedure and were therefore held constant in all optimizations. The biotic ligand-Zn log K had a narrow range (5.3–5.6) in all BLMs summarized in DeForest & Van Genderen (2012), so it was fixed at 5.4. Because we wanted to accommodate a pH effect, but we allowed the model to include both

biotic ligand-H and biotic ligand-ZnOH, we needed to fix one parameter. We chose to fix the log *K* for biotic ligand-H at 6.4 (this was the mean log *K* with values ranging from 5.9 to 6.7, as summarized in DeForest & Van Genderen 2012).

With the constraints specified in Supporting Information, Table S2-1, we were able to update the Zn BLM in a manner consistent with (and analogous to) MLR model development. For species-specific models, the critical accumulation specified in Supporting Information, Table S2-1, was optimized by PEST. For pooled models, where the biotic ligand parameters spanned multiple taxa, the PEST procedure included an additional BLM run, where the BLM was also used in speciation mode to calculate the critical accumulation for each species and effect level at each step of the PEST optimization. The output of the Zn BLM optimizations is summarized in Supporting Information, Table S2-2.

Comparison of TMFs in the toxicity dataset to natural waters

Following the approach described in Brix et al. (2021) for evaluating Cu toxicity datasets, the water chemistries in the toxicity dataset for each species (acute and chronic) were compared with conditions in natural US waters using a principal component analysis (PCA) of the *Z* scores of log(DOC), log(Hard), and pH. This analysis was conducted (prcomp() in R; R Core Team, 2021) to reduce the dataset to two dimensions (axes). The PCA model was used to predict the axis scores for the log-transformed Zn toxicity data (predict(), in R; R Core Team, 2021). A scatterplot of axis 2 versus axis 1 scores of the toxicity dataset for each species was overlain on a scatterplot for the US waters dataset, which was used to provide a qualitative visual assessment of how well the TMF conditions in the toxicity tests represented the US waters dataset.

Model comparison

Multiple linear regression models and the optimized BLM were compared using the metrics described in Garman et al. (2020), Brix et al. (2020, 2021), and Besser et al. (2021). Because, as shown later, clear differences in how TMFs influence Zn toxicity to invertebrates and fish were not found, pooled rather than species-specific models were considered the most relevant for criteria development and were thus used for model comparisons. In addition, model comparison metrics for the USEPA's existing hardness model (USEPA, 1987, 1996) were calculated to evaluate whether the Zn MLR models and the BLM provide an improvement over the existing hardness model. The USEPA's hardness model has a ln(Hard) slope of 0.8473.

To assess model accuracy, scatterplots of observed EC*x* versus predicted EC*x* were developed with diagonal lines denoting 1:1 agreement and factor of ±2 agreement. In the present paper, factors of agreement between observed and predicted EC*x* are always calculated based on untransformed (i.e., not log-transformed) EC*x*. To provide more quantitative assessments of the relationship between predicted and observed values, cumulative empirical frequency distributions of the absolute factors of agreement were plotted. Absolute factors of agreement are calculated for each observation as the larger of the paired observed and predicted EC*x* divided by the smaller of the paired observed and predicted EC*x*; hereafter referred to as absolute prediction ratios (Equation 6). The cumulative empirical frequency distributions of absolute prediction ratios for the MLR model, BLM, and USEPA's hardness model were plotted together for direct comparison along with a "null" model. The null model is the geometric mean of EC*x*s for each species with no bioavailability adjustment.

$$\text{Absolute prediction ratio} = \frac{\max(EC_{x,\text{observed}}, EC_{x,\text{predicted}})}{\min(EC_{x,\text{observed}}, EC_{x,\text{predicted}})} \quad (6)$$

Patterns in the deviations between observed and predicted EC*x* were evaluated and compared across models by plotting the ratio of observed-to-predicted EC*x* versus observed EC*x* and each of the TMFs (DOC, pH, hardness). These data were plotted on log10 scales (except for pH), and linear regressions were performed with log10-transformed ratios and TMFs to test for significant trends associated with individual parameters.

Lastly, the above metrics were integrated into an overall model performance score (MPS) as described in Garman et al. (2020). The MPS comprises six components: (1) the R^2 of the linear regression model of observed versus predicted EC*x*, (2) the percentage of values within a factor of 2.0 ($RF_{x,2.0}$) of their paired observed value, and the slope of observed-to-predicted EC*x* ratios versus (3) observed EC*x*, (4) DOC, (5) hardness, and (6) pH.

One modification was made to the scoring method for model residuals. Garman et al. (2020) suggested calculation of residual scores (RS) based on the following equation:

$$\text{Residual score} = 1 - \frac{2}{\pi}\tan^{-1}(|\text{slope}|) \quad (7)$$

Besser et al. (2021) modified the residual scores by weighting the slope (s) parameter for each TMF by the associated *p* value (*p*) of the regression of the residual versus each independent variable (*i*):

$$RS_i = \frac{2}{\left(1 + 10^{|S_i \times (1-p_i)|}\right)} \quad (8)$$

The MPS was then calculated as:

$$\text{MPS} = \frac{R^2 + RF_{x,2.0} + RS_{\text{pred vs. obs}} + RS_{\text{pH}} + RS_{\text{DOC}} + RS_{\text{Hardness}}}{6} \quad (9)$$

The MPS was computed using the mean score, across species, after application of the pooled model to individual species rather than to the complete pooled dataset. As noted in Brix et al. (2020), the R^2 based on the pooled dataset will be higher because it increases the range of (variance in) the independent

variable due to differences in sensitivity (e.g., EC50s) of the species in the pooled dataset. This higher R^2 would not reflect an ability of the pooled model to better account for the effects of TMFs on metal toxicity for all species. In addition, residuals in the pooled dataset may mask patterns in the residuals for individual species (Brix et al., 2021).

Ambient water quality criteria

Acute and chronic Zn HC5s were developed by following USEPA guidelines for derivation of AWQC (USEPA, 1985) to compare how HC5s based on these two models respond to changing DOC, hardness, and pH conditions. The first step was to develop genus sensitivity distributions (GSDs) based on the compiled acute and chronic toxicity data. Median effect concentrations were used for the acute data and EC20s were used for the chronic data. Pooled species models were used to develop the GSDs, which means that the same model is applied to all species in the GSD, with only the sensitivity terms adjusted for each genus.

For development of HC5 equations, the acute and chronic MLR models were rearranged (Equation 10) and then applied to the acute EC50 and chronic EC20 data to determine the intercept (or sensitivity) of each observation in the dataset.

$$\mathrm{Int}X_i = \ln(\mathrm{ECX}_{\mathrm{meas},i}) - b_1[\ln(\mathrm{DOC}_{\mathrm{meas},i})] - b_2[\ln(\mathrm{Hard}_{\mathrm{meas},i})] - b_3[\mathrm{pH}_{\mathrm{meas},i}] \quad (10)$$

where $\mathrm{Int}X_i$ is the observation-specific intercept (i.e., sensitivity), calculated from the reported ECx, DOC, hardness, and pH of each individual observation (meas,i).

The mean intercepts were then calculated for each species and endpoint, where the intercept for the most sensitive endpoint was retained as the species mean intercept. The species mean intercept is termed the critical intercept, which is equivalent to species mean chronic or species mean acute values. These critical intercepts associated with acute EC50s and chronic EC20s are referred to as CI50s and CI20s, respectively.

Once the critical intercepts were calculated, GSDs were developed based on the genus mean critical intercepts. For the BLM, GSDs were developed based on acute and chronic critical accumulation concentrations on the biotic ligand. To derive these critical accumulations, the Zn BLM was run in "speciation mode" for each of the acute and chronic tests compiled, and the genus mean critical accumulation concentrations were calculated in a manner analogous to the approach used to calculate genus mean critical intercepts. Critical accumulation concentrations associated with acute EC50s and chronic EC20s are referred to as CA50s and CA20s, respectively.

Acute and chronic GSDs were developed for the MLR-based CI50s and CI20s and for the BLM-based CA50s and CA20s. The fifth percentiles of these MLR- and BLM-based acute and chronic GSDs were then calculated following USEPA guidelines (USEPA, 1985). For acute criteria, the fifth percentile of the GSD is divided by 2 to estimate a low effect concentration, whereas chronic criteria are set equal to the fifth percentile of the GSD without any adjustment. Following USEPA (1985) terminology, the fifth percentile of the acute GSD divided by 2 is the acute criterion and is referred to as the criterion maximum concentration. The fifth percentile of the chronic GSD is referred to as the final chronic value (FCV). The FCV is typically the basis for the chronic criterion, which is referred to as the criterion continuous concentration. For the present study, to be clear that we are not referring to USEPA-recommended values, we use the generic HC5 term to describe the fifth percentile of the GSD.

RESULTS

Zn toxicity database

The complete acute and chronic toxicity database is provided in Supporting Information, Table S3-1 in Data S3. The database includes test details and identifies the tests used to develop the MLR models and to optimize the BLM, as well as the tests used to develop the GSDs for HC5 calculations.

MLR models

Six species (*C. dubia*, *D. magna*, *D. pulex*, *O. mykiss*, *P. paludosa*, and *P. promelas*) met minimum DOC, hardness, and pH requirements for acute MLR model development and five species (*C. dubia*, *D. magna*, *L. stagnalis*, *O. mykiss*, and *R. subcapitata*) met the minimum requirements for chronic MLR model development (DOC, hardness, and pH ranges are provided for each exposure type and species in Supporting Information, Table S4-1 in Data S4).

Evaluation of tests with organic buffers. Initial MLR models were evaluated to assess whether any subsets of tests should be eliminated. First, evidence of effects of organic pH buffers, such as 3-*N*-morpholinopropane sulfonic acid, 2-morpholinoethanesulfonic acid, and 4-butanesulfonic acid), on pH-dependent Zn toxicity was assessed. Consistent with a study that specifically evaluated the influence of buffers on Zn toxicity to *D. magna* and *R. subcapitata* (De Schamphelaere et al., 2004), we concluded that organic buffers do not have an important influence on pH-dependent Zn toxicity. For example, the percentage of tests with the predicted ECx being within a factor of 2.0 of observed Ecx was similar for tests with and without organic buffers, indicating that there is not a bias in accuracy (see Supporting Information, Data S5, for details).

Evaluation of life stage effects on sensitivity. The initial MLR models were also investigated for possible life stage or acclimation-related effects on sensitivity, after accounting for bioavailability among tests. This evaluation was particularly important for acute tests where rainbow trout and fathead minnows of various life stages have been tested. This resulted in identification of less sensitive life stages and Zn-acclimated test organisms that were excluded from further MLR and BLM modeling (see Supporting Information, Data S6, for details).

TMF interaction terms. For acute models, BIC-selected models for three of six species did not retain interaction terms, and for the chronic EC20 models, BIC-selected models for two of five species did not retain interaction terms. In contrast, the AIC-selected models retained interaction terms for all six species for the acute model and four of five species for the chronic EC20 model. However, when interactions were retained, improvement was generally negligible in terms of adjusted and predicted R^2. For *C. dubia*, for example, the adjusted R^2 increased from 0.729 to 0.784 when TMF interactions were included, but predicted R^2 decreased from 0.623 to 0.484.

In addition, the interaction models sometimes resulted in Zn toxicity predictions that were inconsistent with expected responses to TMFs. For example, under some conditions, the acute *C. dubia* and *O. mykiss* models predict that Zn EC50s decrease (i.e., toxicity increases) with increasing DOC concentration, even when hardness and pH are constant. It is possible that interactions among TMFs can alter the effect of an individual TMF on Zn bioavailability in ways that may seem contrary to mechanistic expectations, but we observed that the MLR models with interactions had a negligible effect on model residuals compared with models without interactions. Accordingly, the remainder of this evaluation focuses on MLR models without interactions. Summary statistics for models with interactions are provided in Supporting Information, Data S7.

Acute MLR models. The TMFs retained in the six species-specific models varied among species, including for each of the three daphnid species (Table 1). Both the AIC and the BIC retained hardness in five of six species-specific models and DOC in two of six. The BIC retained pH in two of six species-specific models and the AIC retained pH in three. Model performance differed across species. The AIC and the BIC retained no terms in the *P. promelas* model, which had an adjusted R^2 of approximately 0 (Table 1), whereas the *P. paludosa* model, which retained two to three TMFs (BIC, AIC) and was the only model with a positive pH slope, performed well (adj. $R^2 = 0.90$). For other species, adjusted R^2 ranged from 0.45 to 0.73 and the predicted R^2 were not substantially lower than the respective adjusted R^2 (<0.11 decrease), indicating that the models were not overparameterized (Table 1).

For the pooled model, AIC retained all three TMFs whereas BIC retained DOC and hardness (Table 1). Because AIC retained pH in four of six species-specific models, and there is a mechanistic basis for inclusion of pH in the model, we used the AIC-selected model for the remaining evaluations in the present study. As stated in Brix et al. (2020), in addition to assessing statistical results in evaluating and selecting models, it is also important to consider known TMF mechanisms. When the pooled model was applied to each of the six species comprising the model, the adjusted R^2 ranged from −0.21 to 0.71 (Table 1). The substantial decreases in adjusted R^2 when the pooled model is applied to some individual species is not surprising given the variability in the TMF terms retained in each of the species-specific MLR models (the adjusted R^2 for *C. dubia* and *P. paludosa*, for example, decreased by 0.69 and 0.18, respectively). For *D. magna*, *D. pulex*, and *O. mykiss*, however, the adjusted R^2 decreased by ≤0.07 based on the pooled model compared with their species-specific models. Variance inflation factors for the pooled model ranged from 1.04 to 1.05 for the three TMFs, indicating that the model is not overparameterized; VIFs <3 are typically considered acceptable (Zuur et al., 2010).

For the species-specific models, the percentages of predicted EC50s within a factor of 2.0 of observed EC50s ranged from 78% for *P. promelas* to 100% for *P. paludosa* (Supporting Information, Figure S8-1 in Data S8). For *P. promelas*, however, note that no TMFs were retained by AIC or BIC, so the model was a null model. When the pooled model was applied to each of the six species, the percentages of predicted EC50s within a factor of 2.0 of observed ranged from 58% for *C. dubia* to 100% for *P. paludosa* (Figure 1). *Ceriodaphnia dubia* had the greatest decrease in the number of samples with predicted

TABLE 1: Acute Zn multiple linear regression model statistics (no toxicity modifying factor interaction terms)

						Slopes			
Species	Effect level	*n*	AIC/BIC	Adj. R^2	Pred. R^2	ln(DOC)	ln(Hard)	pH	Intercept
C. dubia	EC50	24	AIC, BIC	0.729	0.623	–	0.282	−0.862	10.532
D. magna	EC50	93	AIC, BIC	0.449	0.429	–	0.507	–	4.429
D. pulex	EC50	26	AIC, BIC	0.569	0.511	0.297	0.837	–	3.185
O. mykiss	EC50	77	AIC, BIC	0.501	0.456	–	0.990	−0.392	4.286
P. promelas	EC50	18	AIC, BIC	0.000	−0.121	–	–	–	6.716
P. paludosa	EC50	11	AIC	0.902	0.842	0.233	0.808	0.106	1.698
			BIC	0.897	0.830	0.247	0.827	–	2.364
Pooled (all)	EC50	249	AIC	–[a]	–	0.127	0.600	−0.120	–[b]
			BIC	–[c]	–	0.115	0.574	–	–[d]

[a]Adj. R^2 for individual species based on application of the AIC-selected pooled model: *C. dubia* = −0.04, *D. magna* = 0.42, *D. pulex* = 0.50, *O. mykiss* = 0.49, *P. promelas* = −0.21, and *P. paludosa* = 0.71.
[b]Species-specific intercepts for the AIC-selected pooled model: *C. dubia* = 3.843, *D. magna* = 4.982, *D. pulex* = 4.949, *O. mykiss* = 3.956, *P. promelas* = 4.784, and *P. paludosa* = 4.338.
[c]Adj. R^2 for individual species based on application of the BIC-selected pooled model: *C. dubia* = −0.06, *D. magna* = 0.43, *D. pulex* = 0.52, *O. mykiss* = 0.48, *P. promelas* = −0.13, and *P. paludosa* = 0.87.
[d]Species-specific intercepts for the BIC-pooled model: *C. dubia* = 3.063, *D. magna* = 4.183, *D. pulex* = 4.114, *O. mykiss* = 3.164, *P. promelas* = 4.018, and *P. paludosa* = 3.605.
AIC = Akaike information criterion; BIC = Bayesian information criterion; DOC = dissolved organic carbon; Hard = hardness; EC50 = median effect concentration.

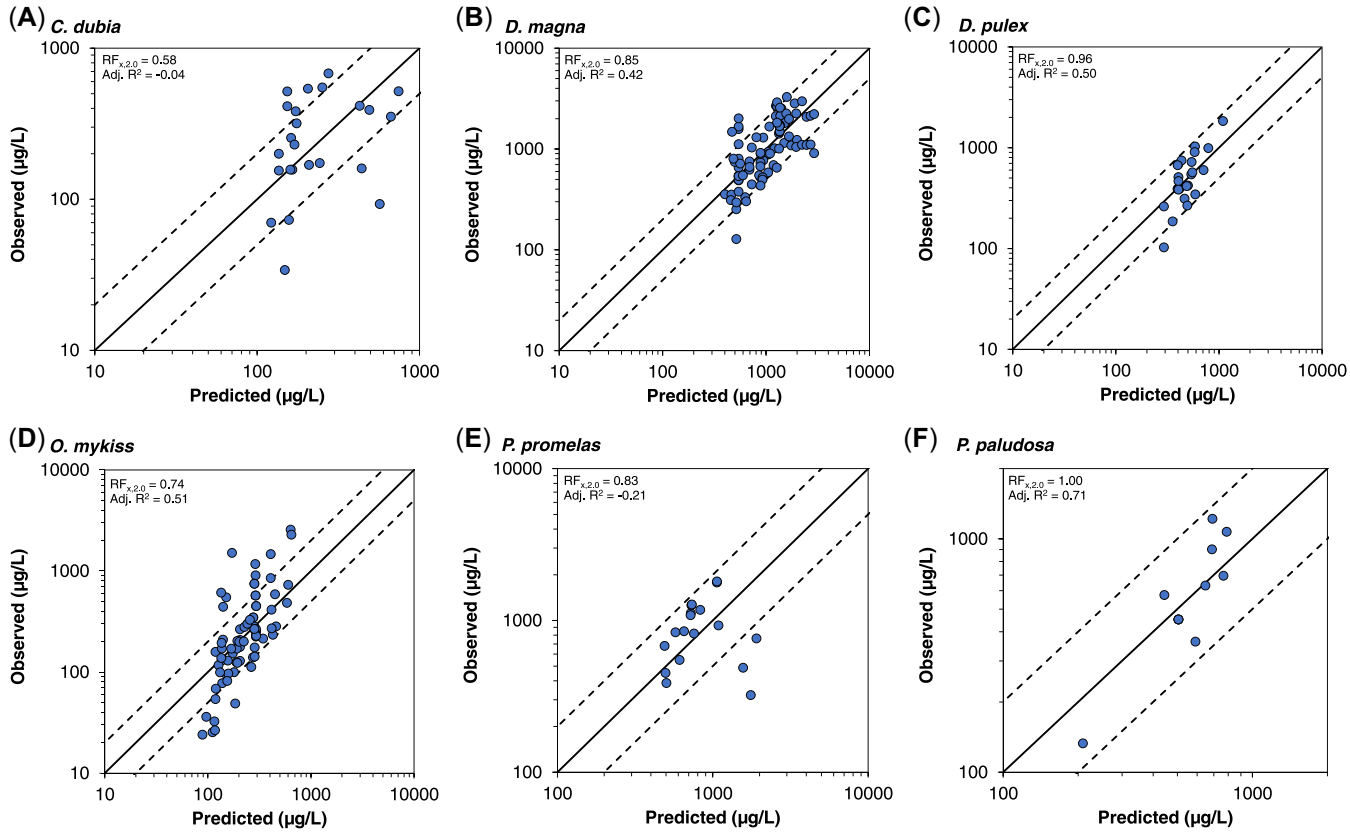

FIGURE 1: Pooled-species model: relationships between observed and predicted acute median effect concentrations (EC50s) for (**A**) *C. dubia* (cladoceran), (**B**) *D. magna* (cladoceran), (**C**) *D. pulex* (cladoceran), (**D**) *O. mykiss* (rainbow trout), (**E**) *P. promelas* (fathead minnow), and (**F**) *P. paludosa* (snail). $RF_{x,2.0}$ = fraction of predicted EC50s within a factor of 2.0 of observed.

EC50s within a factor of 2.0 of observed when the pooled model was applied in place of the species-specific model. The *C. dubia* model has a shallower hardness slope than the pooled model and a steeper pH slope. This results in *C. dubia*-specific and pooled model predictions that increasingly diverge at combinations of low hardness and low pH, and high hardness and high pH.

Chronic MLR models. Chronic MLR models were evaluated based on EC10s, EC20s, and EC50s. In the species-specific models, AIC and BIC retained the same TMFs in the EC10, EC20, and EC50 models (Table 2) and, in general, the adjusted and predicted R^2 were similar within a species for EC10, EC20, and EC50 models (Table 2). The remainder of the chronic MLR results shown in the present study focus on EC20s.

The TMFs retained in the chronic MLR models varied among species. For *C. dubia*, no TMFs were retained based on EC20s. All three TMFs were retained in the *D. magna* and *O. mykiss* models, whereas pH was not retained in the *L. stagnalis* model and hardness was not retained in the *R. subcapitata* model (Table 2). Excluding the *C. dubia* null model, the adjusted R^2 in the species-specific EC20 models ranged from 0.30 for *D. magna* to 0.86 for *L. stagnalis*. The predicted R^2 for *D. magna* and *R. subcapitata* were reduced by less than 0.10 compared with the adjusted R^2, but for *L. stagnalis* and *O. mykiss* the predicted R^2 were reduced by 0.14 and 0.19, respectively, relative to the adjusted R^2 (Table 2). The larger reductions in predicted R^2 for *L. stagnalis* and *O. mykiss* are in part influenced by the smaller sample sizes for these species (7 and 14, respectively) and indicate these models may not be suitable for direct use, but the data are still useful for development of a pooled species model.

For the pooled chronic model comprised of invertebrates and fish, AIC retained all three TMFs in the EC20 model, whereas the BIC-selected EC20 model included only DOC and hardness (Table 2). Because pH was retained in two of the four species-specific models for species in the pooled invertebrate and fish model (*D. magna* and *O. mykiss*), and there is a mechanistic basis for inclusion of pH in the model, we used the AIC-selected pooled chronic model for the remainder of evaluations in the present study (consistent with the decision to apply the AIC-selected pooled acute model in the present study). Similar to the pooled acute model, VIFs were low, ranging from 1.00 to 1.40 for the three TMFs.

We used the AIC model to address the next question, which was whether to adopt the pooled model based on all species (algae, invertebrates, and fish) or the pooled invertebrate and fish model. We found that application of the pooled model based on all species to invertebrates and fish resulted in a large reduction in adjusted R^2 compared with the species-specific models, whereas application of the pooled models based on invertebrates and fish resulted in

TABLE 2: Chronic Zn multiple linear regression model statistics (no toxicity modifying factor interaction terms)

Species	Effect Level	*n*	AIC/BIC	Adj. R^2	Pred. R^2	Slopes			Intercept
						ln(DOC)	ln(Hard)	pH	
C. dubia	EC10	13	AIC, BIC	0.259	−0.088	–	0.489	–	1.851
	EC20	19	AIC, BIC	0.000	−0.114	–	–	–	4.219
	EC50	14	AIC, BIC	0.158	−0.008	–	0.437	–	2.543
D. magna	EC10	47	AIC, BIC	0.117	−0.008	0.373	0.193	−0.238	5.132
	EC20	48	AIC, BIC	0.304	0.215	0.422	0.296	−0.226	4.672
	EC50	50	AIC, BIC	0.453	0.400	0.430	0.399	−0.228	4.477
L. stagnalis	EC10	7	AIC, BIC	0.861	0.733	0.429	0.519	–	3.292
	EC20	7	AIC, BIC	0.856	0.715	0.406	0.488	–	3.631
	EC50	7	AIC, BIC	0.841	0.671	0.368	0.428	–	4.228
O. mykiss	EC10	26	AIC, BIC	0.596	0.518	0.425	0.730	−0.793	7.944
	EC20	14	AIC, BIC	0.563	0.373	0.301	1.093	−0.884	7.627
	EC50	27	AIC, BIC	0.433	0.309	0.311	0.624	−0.481	7.124
R. subcapitata	EC10	37	AIC, BIC	0.629	0.591	0.378	–	−0.992	10.307
	EC20	36	AIC, BIC	0.746	0.704	0.342	–	−1.172	11.950
	EC50	54	AIC, BIC	0.514	0.463	0.209	–	−0.865	10.925
Pooled (all)	EC10	130	AIC, BIC	–[a]	–	0.439	0.294	−0.640	–[c]
	EC20	124	AIC, BIC	–[a]	–	0.376	0.237	−0.646	–[c]
	EC50	152	AIC, BIC	–[a]	–	0.270	0.294	−0.570	–[c]
Pooled (inverts, fish)	EC10	93	AIC, BIC	–[b]	–	0.400	0.351	−0.290	–[d]
	EC20	88	AIC	–[b]	–	0.363	0.327	−0.184	–[d]
			BIC	–[b]	–	0.364	0.270	–	–[d]
	EC50	98	AIC, BIC	–[b]	–	0.327	0.447	−0.232	–[d]

[a]Adj. R^2 for individual species based on application of the pooled all species model: *C. dubia* = −0.29 (EC10), −0.15 (EC20), −0.21 (EC50); *D. magna* = 0.070 (EC10), 0.11 (EC20), 0.20 (EC50); *L. stagnalis* = −0.0075 (EC10), −0.23 (EC20), −0.130 (EC50); *O. mykiss* = 0.43 (EC10), 0.024 (EC20), 0.26 (EC50); *R. subcapitata* = 0.55 (EC10), 0.59 (EC20), 0.33 (EC50).
[b]Adj. R^2 for individual species based on application of the pooled invert+fish model: *C. dubia* = −0.17 (EC10), −0.12 (EC20, AIC), −0.063 (EC20, BIC), 0.009 (EC50); *D. magna* = 0.087 (EC10), 0.29 (EC20, AIC), 0.24 (EC20, BIC), 0.43 (EC50); *L. stagnalis* = 0.57 (EC10), 0.65 (EC20, AIC), 0.83 (EC20, BIC), 0.64 (EC50); *O. mykiss* = 0.52 (EC10), 0.40 (EC20, AIC), 0.44 (EC20, BIC), 0.41 (EC50).
[c]Species-specific intercepts for the pooled models based on all species: *C. dubia* = 7.259 (EC10s), 7.676 (EC20s), 7.374 (EC50s); *D. magna* = 7.454 (EC10s), 7.990 (EC20s), 7.579 (EC50s); *L. stagnalis* = 9.282 (EC10s), 9.824 (EC20s), 9.359 (EC50s); *O. mykiss* = 8.594 (EC10s), 9.228 (EC20s), 9.153 (EC50s); and *R. subcapitata* = 6.676 (EC10s), 7.258 (EC20s), 7.713 (EC50s).
[d]Species-specific intercepts for the pooled models based on invertebrates and fish: *C. dubia* = 4.256 (EC10s), 3.626/2.439 (EC20s; AIC/BIC), 3.907 (EC50s); *D. magna* = 4.745 (EC10s), 4.299/3.243 (EC20s; AIC/BIC), 4.405 (EC50s); *L. stagnalis* = 6.342 (EC10s), 5.841/4.678 (EC20s; AIC/BIC), 5.982 (EC50s); and *O. mykiss* = 5.777 (EC10s), 5.473/4.342 (EC20s; AIC/BIC), 6.026 (EC50s).
AIC = Akaike information criterion; BIC = Bayesian information criterion; DOC = dissolved organic carbon; Hard = hardness; EC10 = 10% effect concentration; EC20 = 20% effect concentration; EC50 = 50% effect concentration.

negligible to modest reduction in adjusted R^2 (Table 3). Consequently, for subsequent evaluations, we selected the pooled invertebrate and fish model to represent invertebrates and fish, and the *R. subcapitata* model to represent algae and plants.

For the species-specific models, the percentages of predicted EC20s within a factor of 2.0 of observed ranged from 78% for *R. subcapitata* to 100% for *L. stagnalis* (Supporting Information, Figure S8-2). When the AIC-selected pooled invertebrate and fish model was applied to the three invertebrate species and one fish species, the percentages of predicted EC20s within a factor of 2.0 of observed ranged from 64% for *O. mykiss* to 100% for *L. stagnalis* (Figure 2).

Cross-validation of MLR models

Cross-validation of the pooled acute model corroborated the stepwise BIC model that retained all three TMFs (DOC, hardness, and pH). The CVs for all three TMF coefficients based on the five training datasets were 13% for DOC, 3.5% for hardness, and 48% for pH (Table 4). The higher CVs for DOC and pH compared with hardness are consistent with observations that DOC and pH have a variable influence on Zn toxicity among species, whereas hardness is more consistent. Similarly, the *p* values for the DOC and pH coefficients for the five training datasets varied more than the hardness coefficient. The *p* values for hardness were always <0.001, whereas for DOC the *p* value ranged from 0.002 to 0.01 and for pH the

TABLE 3: Comparison of adjusted R^2 based on chronic 20% effect concentration species-specific models, pooled invertebrate + fish model, and pooled alga + invertebrate + fish model

Model	*C. dubia*	*D. magna*	*L. stagnalis*	*O. mykiss*	*R. subcapitata*
Species-specific	0.000	0.304	0.856	0.563	0.746
Pooled invertebrate + fish	−0.124	0.289	0.652	0.403	–
Pooled alga + invertebrate + fish	−0.151	0.107	−0.230	0.024	0.574

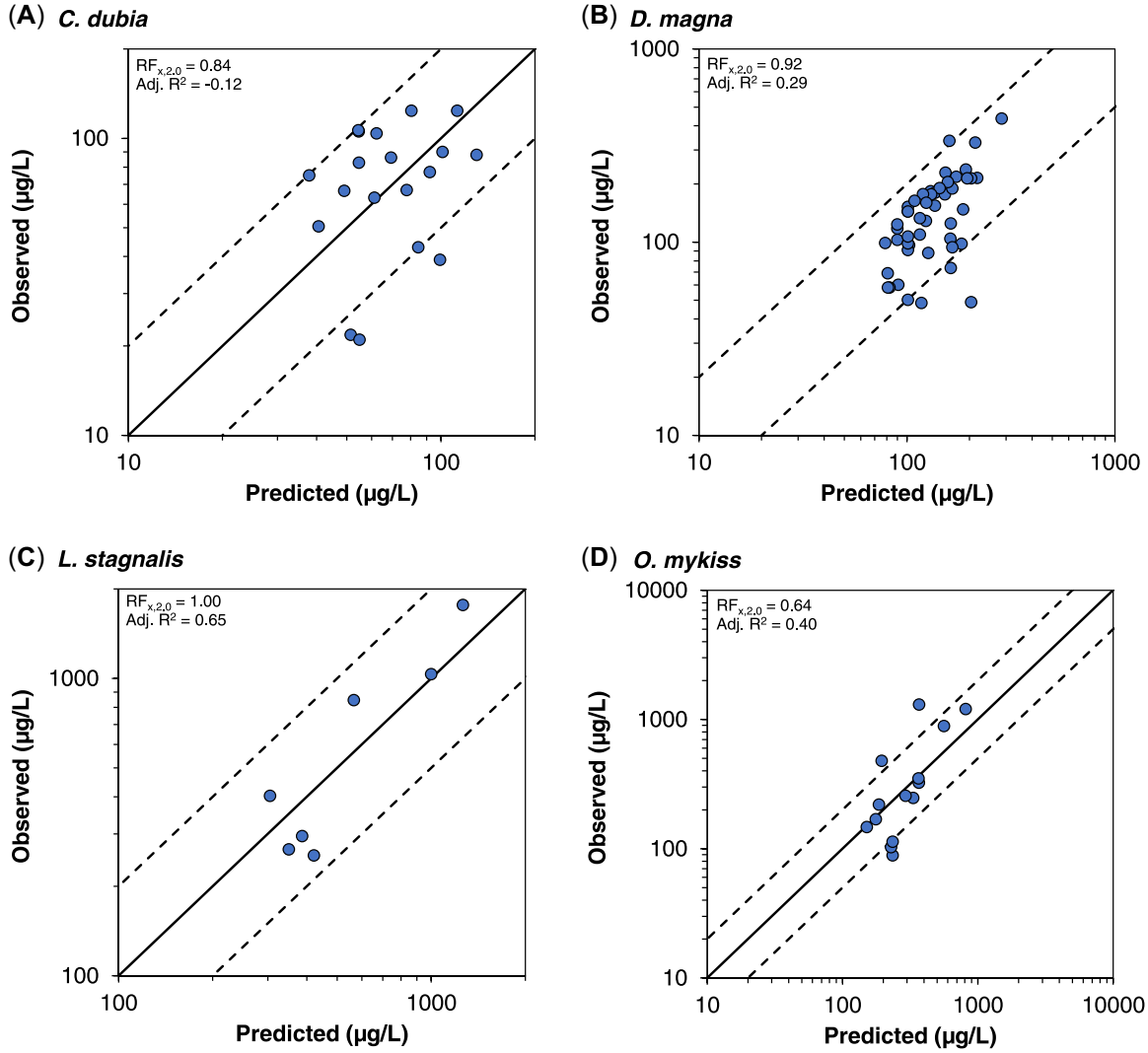

FIGURE 2: Akaike information criterion-selected pooled invertebrate + fish model: relationships between observed and predicted chronic 20% effect concentrations (EC20s) for (**A**) *C. dubia* (cladoceran), (**B**) *D. magna* (cladoceran), (**C**) *L. stagnalis* (snail), and (**D**) *O. mykiss* (rainbow trout). $RF_{x,2.0}$ = fraction of predicted EC20s within a factor of 2.0 of observed.

p value was >0.05 in four of five folds (Table 4). Lastly, for each of the five folds, the RMSE for the training dataset was sometimes less than the RMSE for the test dataset and sometimes greater, but the average RMSEs of 0.60 and 0.62 for the training and test datasets respectively were similar (Table 4). The variability in the RMSE for the training and test datasets for individual folds is likely driven by the higher variability in the DOC and pH coefficients.

In the *k*-fold evaluation of the pooled chronic model, variance in pH coefficients (CV = 49%) was greater across the five folds than for DOC (CV = 13%) and hardness (CV = 11%; Table 4). Ranges in *p* values for TMF coefficients exhibited a similar pattern with *p* values for pH ranging from highly significant to highly nonsignificant (0.003–0.4) whereas DOC and hardness were highly significant for all five folds (<0.001; Table 4). Despite the high CV for the pH coefficient among the five folds in the chronic model, along with highly variable *p* values for the pH coefficient, the overall fit statistics (adjusted R^2 and predicted R^2) for the five cross-validation training runs did not differ from one another by more than 11% (Table 4). This indicates that, despite greater variability in the pH coefficient (in terms of both magnitude and significance), the pooled chronic model, which includes DOC, hardness, and pH, is reflective of the full dataset.

Individual TMF regression analysis

The TMFs retained in the species-specific MLR models varied among species and, in many cases, the slopes for retained TMFs differed by a factor of three or more among species (Tables 1 and 2). The results of univariate regression analyses of individual TMFs demonstrated similar variability among species, with a given TMF having a significant slope for some species but not for others (Supporting Information, Table S9-1 in Data S9). For acute Zn toxicity, the species-specific slopes differed significantly among species for all three TMFs: DOC and hardness ($p < 0.001$) and pH ($p = 0.002$), but

TABLE 4: Selected goodness-of-fit statistics from cross-validation of the pooled model

Model	Adj. R^2	Pred. R^2	AIC	Slope (p value) ln(DOC)	Slope (p value) ln(Hard)	Slope (p value) pH	Training model sample size	Test model sample size	Training RMSE	Test RMSE
Acute										
EC50_k1	0.64	0.62	379.4	0.11 (0.01)	0.60 (<0.001)	−0.13 (0.08)	202	47	0.59	0.67
EC50_k2	0.63	0.61	385.8	0.15 (0.002)	0.63 (<0.001)	−0.12 (0.2)	199	50	0.61	0.60
EC50_k3	0.66	0.63	376.8	0.11 (0.008)	0.62 (<0.001)	−0.033 (0.7)	200	49	0.59	0.67
EC50_k4	0.64	0.62	375.6	0.14 (<0.001)	0.59 (<0.001)	−0.19 (0.02)	199	50	0.59	0.66
EC50_k5	0.63	0.61	394.5	0.12 (0.008)	0.58 (<0.001)	−0.13 (0.1)	196	53	0.63	0.51
Average	0.64	0.62	382.4	0.13	0.60	−0.12	199	50	0.60	0.62
SD	–	–	–	0.017	0.021	0.058	–	–	–	–
CV	–	–	–	0.13	0.035	−0.48	–	–	–	–
Pooled	0.64	0.62	475.1	0.127	0.600	−0.12	–	–	0.60	–
Chronic										
EC20_k1	0.68	0.63	100.1	0.31 (<0.001)	0.38 (<0.001)	−0.32 (0.003)	73	15	0.43	0.67
EC20_k2	0.59	0.54	111.5	0.33 (<0.001)	0.32 (<0.001)	−0.09 (0.4)	70	18	0.48	0.47
EC20_k3	0.69	0.65	109.5	0.39 (<0.001)	0.31 (<0.001)	−0.20 (0.06)	70	18	0.47	0.49
EC20_k4	0.68	0.64	101.5	0.35 (<0.001)	0.29 (<0.001)	−0.11 (0.3)	70	18	0.45	0.57
EC20_k5	0.64	0.60	115.6	0.43 (<0.001)	0.33 (<0.001)	−0.21 (0.06)	69	19	0.50	0.38
Average	0.66	0.62	107.6	0.36	0.33	−0.19	70	18	0.46	0.52
SD	–	–	–	0.05	0.03	0.09	–	–	–	–
CV	–	–	–	0.13	0.11	−0.49	–	–	–	–
Pooled	0.66	0.62	133.0	0.36	0.33	−0.18	–	–	0.47	–

AIC = Akaike information criterion; CV = coefficient of variation; DOC = dissolved organic carbon; RMSE = root mean square error; EC20 = 20% effect concentration; EC50 = 50% effect concentration; k1, k2, k3, k4, k5 = each of five folds in the cross-validation; SD = standard deviation.

not as significantly for hardness ($p = 0.09$). For chronic Zn toxicity, the species-specific slopes differed significantly among species for hardness ($p = 0.002$) and pH ($p = 0.003$), but not for DOC ($p = 0.5$).

Following USEPA guidance for AWQC derivation, a pooled slope would be calculated only when differences among species-specific slopes were not significant, which, for the present study, is DOC for chronic toxicity. The pooled chronic DOC slope in the individual TMF analysis is 0.424 (Supporting Information, Table S9-2), which is comparable to the slope of 0.330 for the pooled chronic MLR model.

Comparison of TMFs in toxicity dataset to natural waters. Understanding how TMFs in the toxicity dataset are distributed relative to natural waters is useful in understanding how well the MLR model developed from the toxicity data will predict toxicity across the full range of natural waters. As described previously, we investigated the overlap of laboratory and natural water concentrations using PCA (Brix et al., 2020, 2021). The first two PCA axes of the natural waters dataset explained 92% of the variance in the dataset (axis 1, 60%; axis 2, 32%). Axis 1 was correlated with pH (–0.70), hardness (–0.64), and DOC (0.32), and axis 2 was correlated with DOC (–0.91) and hardness (–0.42; Supporting Information, Figure S10-1 in Data S10). Similar to the Cu toxicity datasets (Brix et al., 2021), the acute Zn toxicity datasets used to develop the models covered a wide range of conditions, but included a large number of waters with TMF combinations that were outside the range of US waters, and particularly in the upper left quadrant of the plots where DOC concentrations are low (Supporting Information, Figure S10-2). For chronic Zn toxicity, a lower proportion of toxicity tests fell within the upper left quadrant and the data were generally more evenly distributed across the range of natural waters (Supporting Information, Figure S10-3).

Comparison of MLR, BLM, and hardness model performance

For acute toxicity, the total model performance score based on the BLM (0.70) was higher than that based on the pooled acute MLR model (0.68), and both had higher model performance scores than the hardness model (0.63; Table 5). For the individual species, the total model performance scores never varied by more than 0.08 for the MLR model and the BLM.

The R^2 for observed versus predicted acute EC50s based on the MLR model and the BLM were similar for four of the six species, but the R^2 for the BLM was 18% greater for *D. magna* and 15% greater for *O. mykiss* (Table 5). However, the percentages of predicted EC50s within a factor of 2.0 of observed ($RF_{x,2.0}$) were similar for *D. magna* (88% and 82% for the BLM and the MLR model, respectively) and *O. mykiss* (73% for both the BLM and the MLR model). The hardness model had a very low adjusted R^2 for *C. dubia* and *P. promelas*, as did the MLR model and the BLM, a lower adjusted R^2 than the MLR model and the BLM for *D. pulex* and *P. paludosa*, and an R^2 intermediate to the MLR model and the BLM for *D. magna* (Table 5). In general, $RF_{x,2.0}$ was comparable to the MLR model and the BLM (Table 5).

The directionality (i.e., positive, negative) of the slopes of the observed/predicted EC50 ratios versus observed EC50s and TMFs for the MLR model and the BLM were generally consistent (Table 5 and Supporting Information, Figures S12-1–S12-6

TABLE 5: Summary of pooled multiple linear regression model and biotic ligand model performance comparison metrics

					Slopes (*p* values) of observed/predicted EC*x* versus:				
Exposure type	Species	Model	R^2	$RF_{x,2.0}$	Observed EC*x*	pH	DOC	Hardness	Model score
Acute (EC50)	*C. dubia*	MLR	0.10	58%	0.77 (<0.001)	−0.34 (<0.001)	−0.05 (0.657)	−0.49 (0.007)	0.51
		BLM	0.12	42%	0.61 (0.012)	−0.29 (0.004)	−0.13 (0.381)	−0.62 (0.003)	0.49
		Hardness	0.01	42%	0.89 (<0.001)	−0.44 (<0.001)	0.07 (0.652)	−0.75 (<0.001)	0.41
	D. magna	MLR	0.44	82%	0.50 (<0.001)	0.04 (0.255)	−0.04 (0.431)	−0.05 (0.407)	0.77
		BLM	0.62	88%	0.14 (0.054)	0.03 (0.344)	0 (0.941)	−0.18 (<0.001)	0.85
		Hardness	0.46	77%	0.24 (0.008)	0.03 (0.544)	0.11 (0.057)	−0.34 (<0.001)	0.74
	D. pulex	MLR	0.56	96%	0.63 (<0.001)	0.05 (0.676)	0.16 (0.063)	0.19 (0.285)	0.76
		BLM	0.61	92%	0.33 (0.006)	0.08 (0.408)	−0.1 (0.169)	0.16 (0.316)	0.82
		Hardness	0.36	85%	0.59 (<0.001)	0.01 (0.966)	0.3 (<0.001)	−0.1 (0.61)	0.71
	O. mykiss	MLR	0.51	73%	0.66 (<0.001)	0.07 (0.231)	−0.04 (0.703)	0.27 (0.004)	0.70
		BLM	0.66	73%	0.54 (<0.001)	0.1 (0.056)	−0.12 (0.164)	0.26 (0.001)	0.72
		Hardness	0.48	69%	0.49 (<0.001)	−0.11 (0.069)	0.15 (0.145)	−0.05 (0.59)	0.73
	P. promelas	MLR	0.00	83%	0.99 (<0.001)	0.25 (0.171)	0.06 (0.591)	−0.59 (0.003)	0.53
		BLM	0.05	50%	0.65 (0.124)	0.3 (0.204)	−0.33 (0.023)	−0.49 (0.09)	0.48
		Hardness	0.01	61%	1.09 (0.002)	0.28 (0.211)	0.23 (0.1)	−0.9 (<0.001)	0.42
	P. paludosa	MLR	0.80	100%	0.46 (<0.001)	0.13 (0.025)	0.14 (0.055)	0.27 (0.148)	0.79
		BLM	0.69	100%	0.27 (0.128)	0.08 (0.23)	0.01 (0.922)	0.38 (0.041)	0.82
		Hardness	0.63	91%	0.39 (0.035)	0.09 (0.21)	0.25 (<0.001)	0.02 (0.922)	0.79
	All[a]	MLR	–	–	–	–	–	–	0.68
		BLM	–	–	–	–	–	–	0.70
		Hardness	–	–	–	–	–	–	0.63
Chronic (EC20)	*C. dubia*	MLR	0.06	84%	0.83 (<0.001)	−0.03 (0.777)	−0.21 (0.205)	−0.12 (0.437)	0.65
		BLM	0.01	42%	0.91 (0.005)	−0.06 (0.658)	−0.46 (0.035)	−0.2 (0.336)	0.50
		Hardness	0.01	63%	0.89 (0.021)	−0.46 (<0.001)	0.32 (0.218)	−0.8 (<0.001)	0.40
	D. magna	MLR	0.33	90%	0.65 (<0.001)	−0.02 (0.57)	0.07 (0.528)	−0.05 (0.415)	0.75
		BLM	0.47	92%	0.31 (0.006)	0.02 (0.621)	−0.08 (0.456)	−0.05 (0.412)	0.82
		Hardness	0.13	54%	0.38 (0.107)	−0.2 (0.01)	0.71 (<0.001)	−0.67 (<0.001)	0.46
	L. stagnalis	MLR	0.83	100%	0.36 (0.044)	0.21 (0.072)	0.11 (0.485)	0.23 (0.146)	0.82
		BLM	0.73	100%	0.28 (0.21)	0.22 (0.107)	0.04 (0.82)	0.17 (0.37)	0.85
		Hardness	0.68	86%	0.15 (0.6)	−0.05 (0.784)	0.26 (0.192)	−0.18 (0.437)	0.85
	O. mykiss	MLR	0.54	64%	0.61 (<0.001)	0.07 (0.587)	0.05 (0.765)	0.35 (0.11)	0.70
		BLM	0.73	86%	0.28 (0.043)	−0.04 (0.685)	−0.04 (0.734)	0 (0.984)	0.88
		Hardness	0.32	50%	0.56 (0.01)	−0.29 (0.055)	0.38 (0.046)	−0.22 (0.408)	0.57
	All[a]	MLR	–	–	–	–	–	–	0.73
		BLM	–	–	–	–	–	–	0.76
		Hardness	–	–	–	–	–	–	0.57

[a]Average model score for all species.
BLM = biotic ligand model; MLR = multiple linear regression; $RF_{x,2.0}$ = the percentage of values within a factor of 2.0; EC*x* = effect concentration; DOC = dissolved organic carbon.

in Data S12). The directionality of the hardness model slopes for observed/predicted EC50 ratios versus TMFs relative to the MLR model and the BLM varied by species. For example, for rainbow trout, the residuals slopes were negative for pH and hardness based on the hardness model and positive for the MLR model and the BLM, whereas the opposite was observed for DOC (Supporting Information, Figure S12-4). Most conspicuously, the observed/predicted EC50 ratios plotted versus observed EC50s had a positive slope for all three models, which indicates underprediction of acute toxicity in tests with low observed EC50s and overprediction of acute toxicity in tests with high observed EC50s (Table 5 and Supporting Information, Figures S12-1–S12-6).

Total model performance scores based on the pooled chronic MLR model and the BLM were similar (0.73 and 0.76, respectively), but lower for the hardness model (0.57; Table 5). The largest difference between the MLR model and the BLM was for *C. dubia*. Both models had a low R^2 (0.06 and 0.01 for the MLR model and the BLM, respectively), but the $RF_{x,2.0}$ was 84% for the MLR model compared with 42% for the BLM (Table 5). The hardness model likewise had a low R^2 (0.01) and an $RF_{x,2.0}$ intermediate to the MLR model and the BLM (63%; Table 5). Importantly, the null model had an $RF_{x,2.0}$ of 89%, which is greater than all three models (Supporting Information, Figure S12-7). This emphasizes the importance in evaluating multiple performance metrics because the moderately high $RF_{x,2.0}$ for *C. dubia* based on the MLR model might erroneously suggest that the model is better at characterizing the influence of TMFs on Zn toxicity to this species compared with the BLM, whereas the even higher $RF_{x,2.0}$ for the null model indicates that there is actually not a strong TMF signal for the chronic EC20 dataset for *C. dubia*.

As for the acute models, the slopes of the residuals versus observed EC20 for the chronic MLR model and BLM were similar and had the same directionality, whereas the slopes were often the opposite for the hardness model (Table 5 and Supporting Information, Figures S12-7–S12-10). When directionality differed between the MLR model and the BLM, both slopes were near zero. Like the acute models, the relationship between observed/predicted EC20s versus observed EC20s

had a positive slope for all three models and all four species. Again, these positive slopes reflect underprediction of chronic toxicity in tests with low observed EC20s and overprediction of chronic toxicity in tests with high observed EC20s (Table 5 and Supporting Information, Figures S12-7–S12-10).

Ambient water quality criteria

The acute and chronic HC5 equations based on the MLR model are provided in Equations (11) and (12). The TMF slopes in the acute HC5/2 equation are based on the pooled acute EC50 model and the TMF slopes in the chronic HC5 model are based on the pooled chronic EC20 model for invertebrates and fish.

$$\text{Acute HC5/2 (µg/L)} = \exp(2.621 + 0.127 \times \ln(\text{DOC}) + 0.600 \times \ln(\text{Hard}) - 0.120 \times \text{pH}) \quad (11)$$

$$\text{Chronic HC5 (µg/L)} = \exp(3.233 + 0.363 \times \ln(\text{DOC}) + 0.327 \times \ln(\text{Hard}) - 0.184 \times \text{pH}) \quad (12)$$

The fifth percentile acute and chronic critical accumulations on the biotic ligand are 1.042 and 0.1021 nmol/g wet weight, respectively.

Note that the chronic HC5 (Equation 12) and fifth percentile chronic critical accumulations are based on exclusion of the mayfly *Neocloeon triangulifer* in the chronic GSD. The basis for this is discussed in the *Chronic HC5s* section.

Acute HC5s. Comparability in acute HC5s based on the pooled MLR model and the BLM varies among the three TMFs and within individual TMF conditions. For example, the influence of DOC on the acute MLR-based HC5 is more moderate than on the acute BLM-based HC5 (Figure 3A). In addition, because the pooled MLR model selected does not include interaction terms, the relative relationship between MLR-based HC5s and DOC for different pH conditions does not vary, whereas for the BLM the combined influence of DOC and pH is influencing the BLM-based HC5s. The BLM-based HC5s at a pH of 6.5 are greater than the HC5s at a pH of 7.5 and 8.5 when DOC is less than approximately 3 mg/L, but less than the HC5s at a pH of 7.5 and 8.5 when DOC is greater than approximately 3 mg/L (Figure 3A). The BLM results are suggestive of a threshold effect for DOC, where there is a relatively minor effect of DOC on HC5s at low DOC, but DOC has a much greater effect on HC5s because DOC increases above 2–5 mg/L. This threshold appears to be pH dependent.

The MLR- and BLM-based HC5s had the greatest degree of overlap when plotted as a function of hardness (Figure 3B). Similar to the pattern observed for DOC, the relative magnitudes of the BLM-based acute HC5s, when plotted as a function of hardness, varied with pH—the HC5s in waters with a hardness less than approximately 100 mg/L were lower at pH 6.5 than at pH 7.5 and 8.5, but the HC5s were higher at pH 6.5 compared with pH 7.5 and 8.5 as the hardness increased above approximately 140 mg/L (Figure 3B).

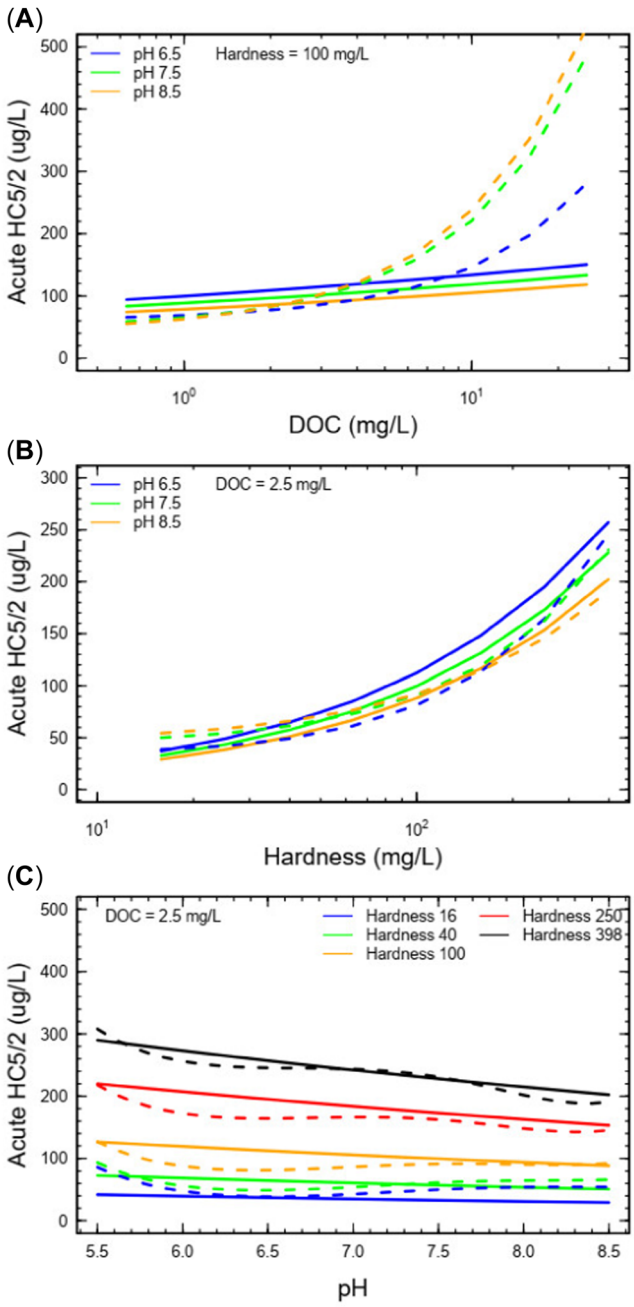

FIGURE 3: Comparison of acute multiple linear regression (MLR)- and biotic ligand model (BLM)-based 5% hazard concentrations (HC5s) as a function of (**A**) dissolved organic carbon (DOC), (**B**) hardness, and (**C**) pH. Solid lines = MLR model; dashed lines = BLM. Toxicity modifying factor ranges are constrained to those in the dataset used to develop the models (see Supporting Information, Table S4-1).

The acute MLR- and BLM-based HC5 generally have high overlap between a pH range of 6.0–8.0 (Figure 3C). The most conspicuous differences between the two models are at a pH of less than approximately 6, where the BLM results in much higher HC5s compared with the MLR model, and at a pH greater than approximately 8, where the BLM-based HC5s begin to increase whereas the MLR-based HC5s continuously decline with increasing pH.

Lastly, uncertainty in the MLR-based HC5s from the pooled acute model was evaluated by evaluating acute HC5s calculated from the models derived from each of the five training datasets. When plotted as a function of DOC (with pH and hardness held at 7.5 and 100 mg/L, respectively), acute HC5s for four of the folds substantially overlapped with the pooled model over a DOC range of <1–30 mg/L (Supporting Information, Figure S11-1A in Data S11). Acute HC5s for one of the folds increasingly diverged from the others with increasing DOC, but was only approximately 40% higher than the pooled model at the highest DOC plotted, which represents a very low percentage of natural waters. When plotted as a function of hardness, acute HC5s for all five folds, as well as the pooled model, substantially overlapped from a hardness of <20–400 mg/L (Supporting Information, Figure S11-1B). Acute HC5s based on four of the folds also show substantial overlap with the pooled model over a pH range of 5.5–8.5, with acute HC5s based on one of the folds diverging relative to the other with decreasing pH (Supporting Information, Figure S11-1C). However, even at the very low pH of 5.5, which does not occur frequently in natural waters, the acute HC5 from the one divergent fold is only approximately 10% less than the acute HC5 from the pooled model. The overall consistency in the acute HC5s among the five folds and the pooled model, especially relative to common ranges of DOC, hardness, and pH conditions in natural waters, provides further evidence that the pooled acute MLR model is not overly sensitive to subsets of the data.

Chronic HC5s. In general, the relative patterns in the MLR- and BLM-based chronic HC5s, as a function of individual TMFs, were similar to those observed for the acute HC5s (Figure 4). However, there is greater overlap in the MLR- and BLM-based HC5s as a function of DOC because the DOC slope in the pooled chronic MLR model (0.363) is greater than that in acute MLR model (0.127).

As indicated in Figure 4, HC5s were calculated from GSDs with and without *N. triangulifer*. Given the available data, *N. triangulifer* is identified as the most sensitive taxon in our dataset. However, data suggest that the recently developed chronic *N. triangulifer* toxicity test may not be fully optimized. For example, 14-day survival was more sensitive than the weight endpoint in Soucek et al. (2020), whereas weight was a substantially more sensitive endpoint than survival in Besser et al. (2021). Based on this observation, we chose to calculate HC5s with and without *N. triangulifer* in the chronic GSD.

One notable difference in the MLR- and BLM-based chronic HC5s is the relative magnitudes of the HC5s, particularly in waters with low DOC, low hardness, or low pH. Interestingly, this difference was more pronounced when the mayfly *N. triangulifer* was included in the GSD—see panels A–C in Figure 4 for HC5s when *N. triangulifer* was included and panels D–F when *N. triangulifer* was excluded. This observation is driven by relatively low EC20s for *N. triangulifer* even in test waters where DOC is high. For example, Besser et al. (2021) reported Zn EC20s of 25, 45, and 55 µg/L in waters with DOC concentrations of 3.9, 29, and 40 mg/L, respectively. The lower BLM-based HC5 for waters with low DOC is due to the stronger influence of DOC on Zn bioavailability in the BLM compared with the relatively shallow DOC slope in the pooled MLR model. It is beyond the scope of the present study, but further evaluations on how test study designs may affect the influence of TMFs on Zn toxicity may help elucidate some of the variability observed among species.

As for the acute HC5s, uncertainty in the MLR-based HC5s from the pooled chronic model was evaluated by evaluating chronic HC5s calculated from the models derived from each of the five training datasets. When plotted as a function of DOC (with pH and hardness held at 7.5 and 100 mg/L, respectively), chronic HC5s based on the five models differed little at low to intermediate DOC concentrations, with greater divergence from intermediate to high DOC. At a DOC of 20 mg/L, for example, the HC5s from the *k*-fold models ranged from approximately 75 to 110 µg/L, compared with an HC5 of 85 µg/L based on the pooled model (Supporting Information, Figure S11-2A in Data S11). When plotted as a function of hardness, chronic HC5s based on the five models overlapped and were consistent with the pooled model over a range of low to very high hardness conditions (Supporting Information, Figure S11-2B). Lastly, when plotted as a function of pH, the chronic HC5s based on the five models overlapped at basic and circumneutral pH, but increasingly diverged with decreasing pH (Supporting Information, Figure S11-2C). However, even at very low pH (e.g., 6.0), the chronic HC5s based on the five models ranged from approximately 32–64 µg Zn/L compared with 52 µg Zn/L for the pooled model. The overall consistency in the chronic HC5s among the five folds and comparison with the pooled model, coupled with the cross-validation metrics summarized above, indicate that the pooled chronic model is not overly sensitive to subsets of the data.

DISCUSSION

k-fold cross-validation of MLR models

A *k*-fold cross-validation is frequently used in machine learning applications to evaluate the ability of a model to fit future data, both to ensure that machine learning models, often developed using multiple potential input parameters, do not over-fit the data and to estimate the true error in predicting future data. In the present study, as well as in other recent MLR model studies for other metals (Brix et al., 2021; Croteau et al., 2021; DeForest et al., 2020), we used predicted R^2, also called "leave one out" cross-validation (a form of *k*-fold cross-validation where $k = n$) to provide a straightforward and easy to understand measure (percentage of variance explained) of how a model fit to all the data would generalize to other data. We have also used AIC and BIC stepwise procedures to help determine which TMFs contribute most to a well-fitting model.

In the present study, we explored the use of *k*-fold cross-validation, with $k = 5$, as an additional, supportive quantitative tool for evaluating the fit and predictive ability of the final acute and chronic pooled MLR models. Although R^2 values and theoretical confidence intervals around model parameters and predicted values provide quantitative information about the

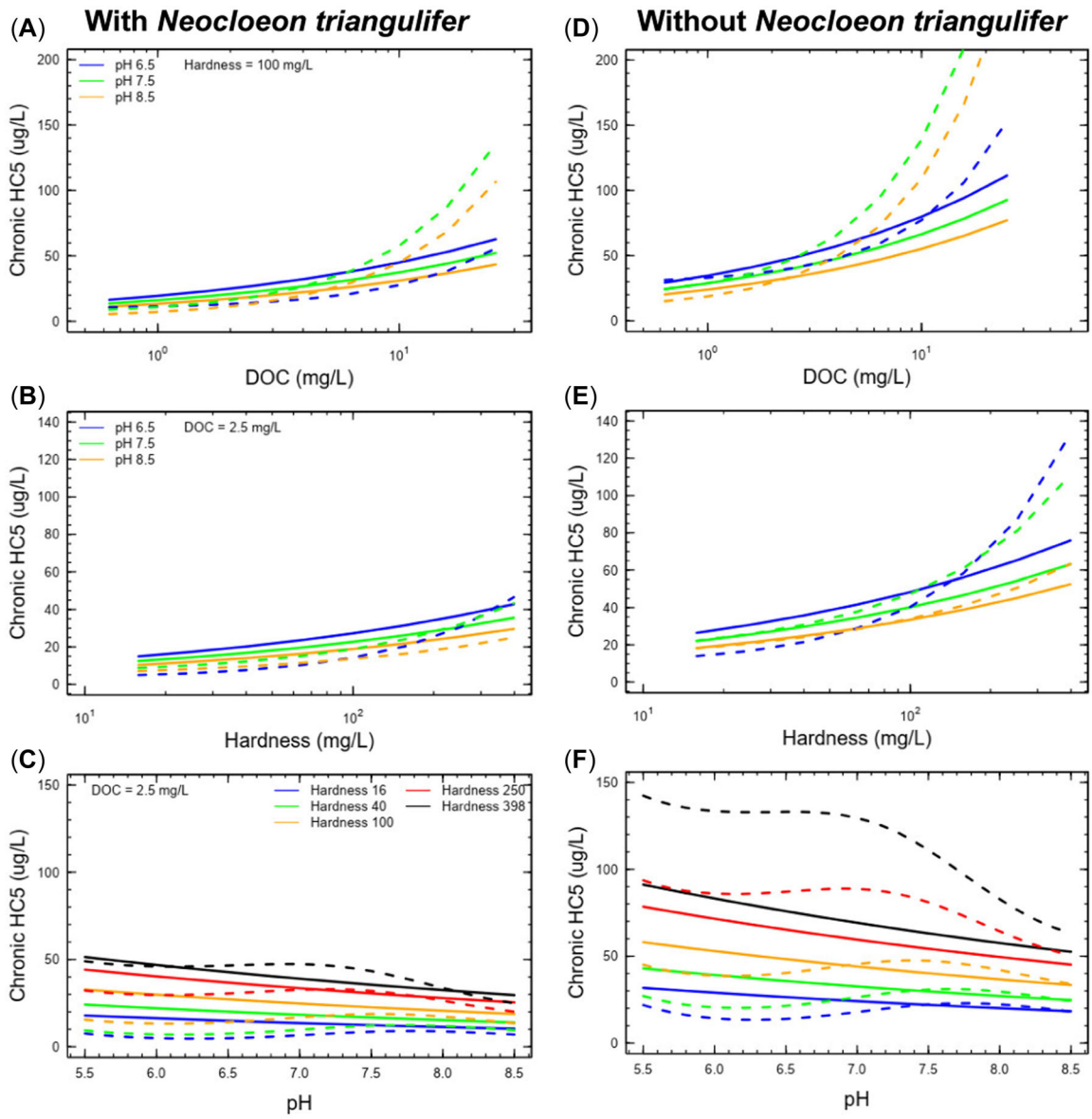

FIGURE 4: Comparison of chronic multiple linear regression (MLR)- and biotic ligand model (BLM)-based 5% hazard concentrations (HC5s): (**A**) dissolved organic carbon (DOC), (**B**) hardness, and (**C**) pH with the mayfly *N. triangulifer* included in the genus sensitivity distribution (GSD), and (**D**) DOC, (**E**) hardness, and (**F**) pH with the mayfly *N. triangulifer* excluded from the GSD. Solid lines = MLR model; dashed lines = BLM. Toxicity modifying factor ranges are constrained to those in the dataset used to develop the models (see Supporting Information, Table S4-1).

potential range of a model's true parameters and how much predicted values might differ from observed values, resampling protocols such as bootstrapping and *k*-fold cross-validation provide concrete examples (via multiple runs or iterations) of the possible range of parameter estimates and fit statistics that are free from theoretical assumptions.

When validation methods use only a single set of cross-validation data, the predicted values of the test dataset are computed using the model developed using the training data and different measures of fit can then be compared to determine whether the training model seems to fit the test data adequately. Depending on the findings, the test data can continue to be held apart from the training dataset and the model based on only the training data used, or the test data can be folded back into the training dataset and a final, full-dataset model created.

A *k*-fold analysis allows for retention of all data in the final model, and also provides *k* examples of how model fit and model parameter coefficients vary for different training-test pairs of data. Although, as for most validation procedures, no absolute criteria exist for deciding how much variability in these *k* measures is too much, the ability to use best professional judgement to evaluate empirical variances in model parameters and fit is valuable. Though in machine learning applications only the mean test RMSE is often retained to provide an estimate of the true RMSE of the final model using all the data, we found that investigation of the variance of coefficient values and fit statistics in the *k* test and training subsets provided additional, useful measures for evaluating assumptions about how TMFs affect toxicity and helped to corroborate findings from analyses we conducted. Finally, application of the training

models through to derivation of HC5s provided a quantitative translation of what this variance means with respect to deriving potential AWQC. We recommend this approach be used in future MLR model development and it may be particularly useful for metals, such as Al, where sample sizes are too small to allow for setting aside independent validation data sets (DeForest et al., 2020).

Individual TMF regression analysis

The analyses of covariance comparing species-specific slopes for individual TMFs was conducted to assess the possibility that MLR models did not adequately capture the influence of a TMF on the Zn EC*x* due to the TMF being underrepresented in the dataset. As in the MLR models, the slopes of TMFs in univariate models differed among species. Pooled TMF regression slopes could only be developed for hardness in acute tests and DOC in chronic tests; for DOC and pH in acute tests and hardness and pH in chronic tests, slopes differed significantly among species ($p < 0.05$). The pooled acute slope for hardness from the individual TMF analysis (0.616) was similar in magnitude to the slope for hardness in the pooled acute MLR model (0.600), especially considering biological variability in toxicity tests and that the datasets varied, and the pooled chronic slope for DOC from the individual TMF analysis (0.424) was likewise similar in magnitude to the slope for DOC in the pooled chronic MLR model (0.330). The individual TMF regression analysis, therefore, did not provide any evidence that the MLR models were not adequately capturing the influence of TMFs.

Overall, the individual TMF regression analysis may be useful for evaluating the MLR model results, but we do not recommend this approach for assembling a multiple TMF model. For the Zn dataset, sample sizes were limited for some species and TMFs, which contributes to greater uncertainty in TMF slopes for those species. This in turn increases the likelihood of observing statistically significant differences in slopes among species, which may preclude development of pooled slopes. Another important factor is that a series of tests in which one TMF is varied may involve different conditions for the other two TMFs. For example, in one DOC series hardness and pH may be low and in another DOC series hardness and pH may be high. The possible influence of how hardness and pH influence the relationship between DOC and Zn EC*x* cannot be accounted for in the univariate, pooled slope evaluation for DOC; however, the MLR model, even without explicitly accounting for TMF interactions, considers each TMF's effect while controlling for the effects of other TMFs in the model.

Model performance scores

Model performance scores based on the method described in Garman et al. (2020), including revision to the slope scoring approach for residuals following Besser et al. (2021), represent a useful way to compare the performance of key model metrics.

In comparing MLR, BLM, and hardness model scores, it is relevant to also consider the null model because the null model characterizes the baseline variability in the toxicity data. Although scores based on the adjusted R^2 and residuals are not relevant to the null model, considering $RF_{x,2.0}$ for the null model can be informative. For example, $RF_{x,2.0}$ based on application of the pooled chronic EC20 MLR model, BLM, and hardness model to *C. dubia* were 84%, 42%, and 63%, respectively, indicating that the MLR model outperformed the BLM and hardness model for this species and metric, but $RF_{x,2.0}$ was even higher (89%) for the null model (Supporting Information, Figure S12-7). An $RF_{x,2.0}$ of 89% for the null model reflects that the observed EC20s do not vary much in response to TMFs (this is reflected in the evaluation of a chronic *C. dubia*-specific EC20 model, in which neither AIC nor BIC retained any TMFs; Table 1). For comparison, application of the pooled acute model to *D. magna* resulted in $RF_{x,2.0}$ of 82%, 88%, and 77% for the MLR model, the BLM, and the hardness model, respectively, compared with 68% for the null model (Supporting Information, Figure S12-2). This indicates that the consideration of TMFs in the *D. magna* models improves toxicity predictions. However, the $RF_{x,2.0}$ of 82% in this acute *D. magna* example is slightly less than the $RF_{x,2.0}$ of 84% based on chronic *C. dubia* example, which was outperformed by the null model and yet both models would effectively receive the same score for $RF_{x,2.0}$.

Garman et al. (2020) recommended a useful set of key metrics for developing these performance scores, and it is perhaps not surprising that subsequent evaluations have identified approaches for revising aspects of these metrics. In addition to modified approaches for calculating the slope scores for residuals, as recommended in Brix et al. (2021) and Besser et al. (2021), the latter of which was applied in the present study, it may be appropriate to further consider how residuals scores are weighted as part of the total performance score (i.e., residuals scores may be weighted too heavily relative to R^2 and $RF_{x,2.0}$). Furthermore, as noted above, it can also be important to consider the $RF_{x,2.0}$ metric relative to the null model to facilitate interpretation of performance metrics for the MLR model and the BLM. Regardless of whether the performance scoring approach is modified, the evaluations from this and recent studies with other metals (Brix et al., 2021; Croteau et al., 2021) indicate that performance metric results should be critically reviewed relative to the questions being asked in the model comparison.

Nonlinear response of BLM to TMFs versus log-linear MLR models

The relationships between metal EC*x* and TMFs may be inherently nonlinear (Brix et al., 2020). A variety of mechanisms can contribute to this nonlinearity. For example, the influence of DOC on EC*x* may be greater at low than high metal concentrations, making DOC a more important TMF for a sensitive species (low EC*x*) versus an insensitive species (high EC*x*) or in chronic exposures versus acute exposures. In addition, the

interactive effects of two or more TMFs can change the slope of a TMF's apparent effect across its range. A good example of this occurs with Al because pH has a strong influence on Al speciation, and the influence of DOC and hardness on Al bioavailability is highly dependent on the Al species present. The inclusion of TMF interaction terms in MLR models can help account for these nonlinearities (DeForest et al., 2018, 2020); although the slope term for a TMF interaction (e.g., DOC and pH) is included as a linear (or log-linear) contribution to the model, the model predictions are nonlinear when one of these TMFs is held constant and the other is varied.

In the present study, we evaluated MLR models with all two-way interactions of the three TMFs, but we did not see improvements in performance relative to models without TMF interactions, and in some cases ECxs responded to a TMF in a manner inconsistent with mechanistic understanding. The inability to identify useful interactive effects in the TMFs, which could replicate the nonlinearities captured in the BLM, may reflect limitations in the Zn dataset. As noted in Brix et al. (2020), developing and testing nonlinear models often requires data over a broad range of TMF conditions, and a linear model often provides a reasonable approximation. For the BLM, however, nonlinear responses are observed. As shown in the HC5 comparisons, the magnitudes of HC5s become greater with increasing DOC and hardness concentrations, and nonlinearity in HC5s was even more pronounced for pH, with HC5s both decreasing and increasing over a pH range of 5.5–8.5 (Figures 3 and 4). In general, though, BLM-based HC5s would vary by less than 1.5-fold from a linear function fit to a given series of BLM-based HC5 plotted against pH (e.g., Figures 3C and 4F), which is in the range of biological variability and indicative that a linear approximation of the HC5 as a function of pH is reasonable (Erickson et al., 1987, for example, reported a range of greater than 6 for replicate Cu LC50s for *P. promelas*).

The pattern of acute HC5s being higher at low pH (<6) and then leveling from a pH of 6 to 8.5 is consistent with Santore et al. (2002), who showed this same relationship for acute Zn toxicity to rainbow trout. They attributed this pattern to a combination of complexation and competition. The competitive effects include displacement of Zn at the biotic ligand due to proton competition in low pH conditions. Protons could also displace Zn from DOC, but the net effect was reduced Zn bioavailability because the tests they evaluated had low DOC. Then, at high pH, Zn hydroxide complexes become more important, thereby reducing Zn ion activity and binding to the biotic ligand. The acute pooled MLR model, without TMF interactions, predicts decreasing EC50s with increasing pH (as reflected in the acute HC5s in Figure 3). Overall, this tends to result in lower HC5s based on the MLR model compared with the BLM under low pH (<6) conditions. The inability of the MLR model to detect a greater reduction in Zn toxicity at pH <6 is likely driven by the general lack of toxicity tests in the MLR dataset conducted at pH conditions <6.

A somewhat similar relationship between HC5s and pH was observed for chronic toxicity based on the BLM, particularly the decreasing relationship between low and circumneutral pH (Figure 4). However, HC5s tend to increase from pH 7 to pH 8, and then begin to decrease at pH >8. The primary difference between the Zn BLM applied to the acute data and the chronic data is the optimization using PEST, which resulted in different acute and chronic log *K*s for ZnOH, Mg, and Na on the biotic ligand. The difference in log *K*s for acute and chronic Zn toxicity results in different simulated HC5 patterns when both the pooled acute and chronic BLMs are applied across a range of TMF conditions. Although both sets of pooled BLM parameters share binding constants for Zn^{2+} and H^{+}, the pooled chronic BLM was optimized with a higher $ZnOH^{+}$ binding constant at the biotic ligand. This explains the slight differences in simulated pH effects when comparing the acute and chronic BLMs. In contrast, the pooled acute and chronic MLR models have similar pH slopes of −0.120 and −0.184, respectively, so both acute and chronic HC5s respond comparably. Overall, there is relatively high overlap in the MLR- and BLM-based HC5s over the pH range of 6–8.5 (when *N. triangulifer* is excluded from the species sensitivity distribution), although there is greater divergence for low hardness waters between pH 6 and 7 (Figure 4F).

Observed/predicted ECx ratio versus observed ECx

For almost all species modeled in both acute and chronic tests, there is a statistically significant ($p < 0.05$) positive slope for the relationship between the observed/predicted EC*x* ratio versus observed EC*x* (Supporting Information, Data S12). This was observed for both the MLR models, the BLM, and hardness model. The positive slope means that EC*x* is overestimated when the observed EC*x* is low and underestimated when the observed EC*x* is high.

The basis for the positive slope is due to greater variability in observed EC*x* than in predicted EC*x*. For the pooled acute MLR model, the predicted EC50s for each species used to develop the model range from 4- to 7-fold, whereas observed EC50s range from 6- to 107-fold. Similarly, for the pooled chronic MLR model, the predicted EC20s for each species used to develop the model range from 3- to 5-fold, whereas observed EC20s range from 6- to 15-fold. In the observed EC*x* versus predicted EC*x* (i.e., 1:1) plots, this results in a higher proportion of predicted EC*x* falling below the 1:1 line when observed EC*x* is lower and a higher proportion falling above the 1:1 line when observed EC*x* is higher. Less variability in predicted EC*x* compared with observed EC*x* suggests that the models are underestimating the influence of known or unknown TMFs. If so, it is interesting that that this pattern was observed for both the MLR and hardness models and the BLM, which were developed using quite different approaches.

To evaluate the issue of positive slopes between observed/predicted EC*x* ratios and observed EC*x* further, we focused on application of the pooled acute MLR model and BLM to acute Zn toxicity to *D. magna*. The acute *D. magna* dataset is large ($n = 93$) and comprises data from 10 separate studies (representing six different research laboratories). The slopes of the relationships between observed/predicted EC50 ratios and observed EC50 were 0.138 for the BLM and 0.500 for the MLR

model (Supporting Information, Figure S12-2). However, slopes of observed/predicted EC50 ratios versus DOC, hardness, and pH were low for both models and insignificant ($p > 0.05$; Supporting Information, Figure S12-2). As such, the magnitudes of the observed/predicted EC50 ratios versus observed EC50 slopes are perhaps unexpected. Thus, we next focused on the lowest observed EC50s for which the EC50 is always overpredicted (i.e., toxicity underpredicted) and the highest observed EC50s for which the EC50 is always underpredicted (i.e., toxicity overpredicted) to further evaluate this observation.

Of the 23 lowest observed acute *D. magna* EC50s, all have an observed/predicted EC50 ratio less than one. There is no variable that is clearly unique to these 23 tests. For example, the 23 tests are from five separate studies, with a DOC range of 0.28–5.4 mg/L, a hardness range of 14–105 mg/L, and a pH range of 6.0–8.3. On the other end of the observed EC50 range, the 10 highest observed EC50s all have an observed/predicted ratio greater than one, are likewise composed of multiple studies (four), and reflect a large DOC range of 0.3–17.3 mg/L and pH range of 6.8–8.2. These 10 tests are all characterized by moderate to high hardness, with a range of 122–525 mg/L. There was a total of 45 acute tests conducted between this hardness range, with 30 having a predicted/observed EC50 ratio greater than one and 15 having a ratio less than one. Thus, although there was a tendency to underpredict the EC50 for higher hardness test waters, there was not a consistent bias in these tests.

Although we cannot entirely rule out the possibility, based on this exploration of the acute *D. magna* dataset, it does not appear that the positive slope between the observed/predicted EC*x* ratio and observed EC*x* is due to inadequately characterizing the influence of TMFs. It should also be noted that both the pooled acute MLR model and the acute *D. magna*-specific MLR model had similar positive slopes for observed/predicted EC50 versus observed EC50 (0.500 for the pooled model and 0.545 for the *D. magna*-specific model), so the observation is not just an artifact of applying the pooled model to *D. magna*. Lastly, again, given that the same pattern is observed for the BLM, this suggests that the observation is not just driven by statistical limitations with the dataset. It may be that other biological variables are contributing to the patterns observed, but this is still uncertain. An important remaining element to consider for the present study, therefore, is the adequacy of MLR models and BLMs for Zn water quality criteria development.

MLR and BLM considerations for ambient water quality criteria

The goal of the present study was to develop Zn MLR models and update Zn BLMs that could be compared for performance and considered for development of bioavailability-based AWQC for Zn. We found that there was high variability in how TMFs influenced Zn toxicity among species, in terms of both the TMFs retained in the species-specific models and the TMF slopes, and that there were often patterns in residuals for both the MLR model and the BLM. However, there was also general consistency in how both models responded to TMFs and the performance scores were comparable between models. From an AWQC perspective, perhaps the most important question is whether the use of either model can improve accuracy in meeting the targeted level of protection, especially compared with existing hardness models.

A common metric that has been historically applied for evaluating the performance of BLMs, and subsequently MLR models, is whether there is a factor of two agreement between observed and predicted EC*x* (i.e., $RF_{x,2.0}$; Santore et al., 2001). As previously discussed, $RF_{x,2.0}$ can also be calculated for a null model. Based on a synthesis of data by exposure type and effect concentrations (i.e., acute EC50s and chronic EC20s), total $RF_{x,2.0}$ values are summarized in Table 6 for the MLR model, the BLM, the hardness model, and the null model. As shown, total $RF_{x,2.0}$ is similar between the MLR model and the BLM, with a difference of 3% and 4% points for acute EC50s and chronic EC20s, respectively. Total $RF_{x,2.0}$ for the hardness model was lower than both the MLR model and the BLM, especially for chronic EC20s ($RF_{x,2.0}$ of 58% compared with 83% and 86% for the BLM and the MLR model, respectively).The null model values were 9%–15% less than those based on the model with the highest value.

Although total $RF_{x,2.0}$ for the hardness and null models are less than both the MLR model and the BLM, the difference is generally small (with a greater reduction in $RF_{x,2.0}$ based on application of the pooled model to chronic EC20s). At a high level, this suggests that both the MLR model and the BLM are an improvement over the hardness and null model, but not substantially. To evaluate this further, absolute deviations from an observed/predicted EC*x* ratio of 1 were calculated for the acute EC50 dataset and the chronic EC20 dataset for the MLR model, the BLM, and the hardness model (for simplicity, the null model was excluded because it does not currently provide a basis for existing Zn criteria). The absolute deviations were plotted versus DOC, hardness, and pH (see panels A–C in Figure 5 for acute data and panels D–F in Figure 5 for chronic data). As shown by the trendlines, the hardness model tends to have greater deviations from 1 than the MLR model and the BLM over most TMF conditions in the chronic dataset, and less conspicuous deviations in the acute dataset. This is not unexpected because the hardness model had a relatively lower total $RF_{x,2.0}$ score for the chronic dataset compared with the acute dataset (Table 6).

For acute toxicity, the relationship between absolute deviations and both DOC and pH is variable among the three model types, but a similar pattern is observed for hardness. For

TABLE 6: Factor of two ($RF_{x,2.0}$) agreement for multiple linear regression models, biotic ligand models, hardness model, and the null model for invertebrates and fish

		$RF_{x,2.0}$			
Exposure type	Effect level	MLR	BLM	Hardness	Null
Acute	EC50	81%	77%	73%	69%
Chronic	EC20	86%	83%	58%	78%

$RF_{x,2.0}$ = the percentage of values within a factor of 2.0; BLM = biotic ligand model; MLR = multiple linear regression; EC50 = median effect concentration.

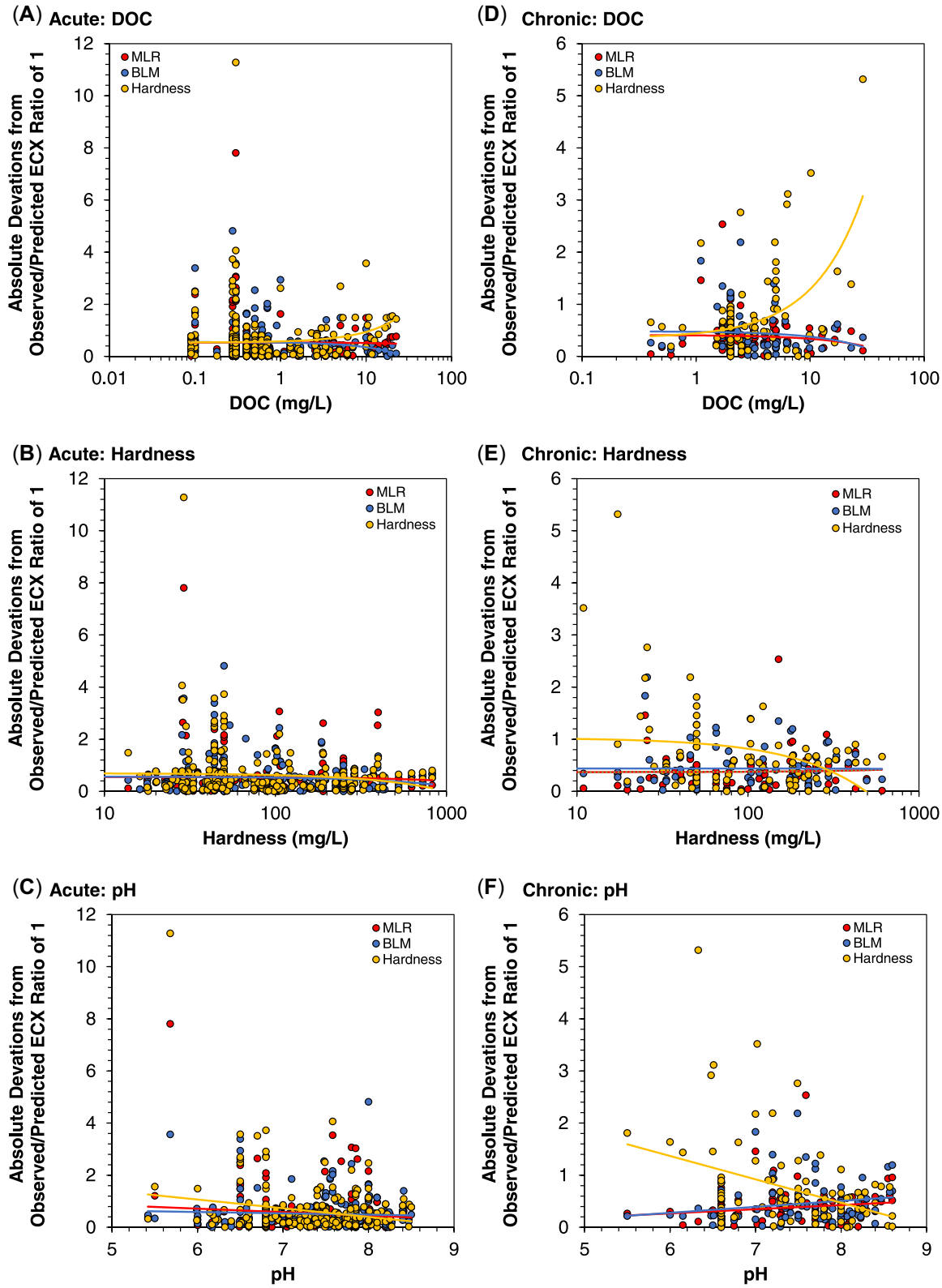

FIGURE 5: Absolute deviations from observed/predicted effect concentration (EC) ratio of 1 based on application of multiple linear regression, biotic ligand model, and hardness models to all acute EC50 data (panels **A–C**) and all chronic EC20 data (panels **D–F**). Linear trend lines help compare relative deviations among models and highlight where models diverge as a function of toxicity modifying factor. DOC = dissolved organic carbon; MLR = multiple linear regression; BLM = biotic ligand model.

DOC, the absolute deviation trendline for the hardness model is relatively flat at low DOC and then begins to increase at a DOC of approximately 3 mg/L, whereas for the MLR model and the BLM the absolute deviation trendline begins to decrease slightly with increasing DOC (Figure 5A). These patterns indicate that the hardness model, which of course does not account for DOC as a mitigating factor on Zn toxicity, provides poorer toxicity predictions with increasing DOC, whereas the MLR model and the BLM better account for the mitigating influence of increasing DOC on Zn toxicity For pH, the absolute deviation trendline for the hardness model demonstrates reduced performance compared with the MLR model and the BLM at pH < 7, and then strong overlap of all three models over a pH range of 7–8.5 (Figure 5C). Although pH and hardness tend to be correlated in surface waters, the explicit inclusion of pH in the MLR model and the BLM appears to improve model performance in acidic waters. Lastly, for hardness, there was substantial overlap in the absolute deviation trends for all three models. This reflects the relatively strong influence of hardness on acute Zn toxicity, which is captured by all three models.

For chronic toxicity, the hardness model performs more poorly than the MLR model and the BLM for all three TMFs (Figure 5D–F). The MLR model and the BLM have similar trendlines for all three TMFs. Most conspicuous are the MLR model and BLM trendlines for DOC, which show a decreasing trend at moderate to high DOC (Figure 5D). This pattern, meaning that model accuracy is increasing at higher DOC, is similar to that observed for the BLM and DOC in the acute dataset. Although this pattern was not as distinct for the MLR model in the acute dataset, the pattern in the chronic dataset can be attributed to the higher DOC slope in the chronic MLR model than in the acute MLR model (0.363 and 0.127, respectively). The absolute deviation trendlines for the MLR model and the BLM versus hardness are both relatively flat over the entire hardness range (Figure 5E), and the trendlines for both slightly increase with increasing pH (Figure 5F). In contrast, the absolute deviations for the hardness model are greater than those for the MLR model and the BLM up to a hardness of 300 mg/L (Figure 5E) and pH 8 (Figure 5F). Thus, the MLR model and the BLM both perform better than the hardness model in waters over a wide range of TMF conditions.

In summary, although there are significant positive slopes between the observed/predicted ECx ratio and observed ECx based on both the MLR model and the BLM, both models predict observed toxicity values within a factor of 2 for a high percentage of samples: 81% and 77% of acute EC50s for the MLR model and the BLM, respectively, and 86% and 83% of chronic EC20s (Table 6). These percentages are higher than the hardness model, which predicted 73% of acute EC50s within a factor of 2 observed and 58% of chronic EC20s (Table 6). Although the MLR model and the BLM, on average, provide improvement in $RF_{x,2.0}$, they tend to perform more strongly under increasing TMF conditions. This is most apparent for the BLM-based acute EC50 predictions for DOC and hardness, and both the MLR- and BLM-based chronic EC20 predictions for DOC.

CONCLUSIONS

Both the MLR model and the BLM respond to TMFs comparably and have similar model performance scores, and both perform better than the hardness model. The MLR model and the BLM reflect mechanistic understandings of how DOC, hardness, and pH influence Zn bioavailability to freshwater organisms. Variability in both models, including low R^2 for some species, may be due to the relatively small range in the TMF response relative to test variability. The two models perform relatively consistently across species, indicating that additional terms, beyond DOC, hardness, and pH, included in the BLM do not help explain the differences among species. Nevertheless, both the MLR model and the BLM provide improved Zn toxicity predictions relative to the hardness model and therefore can be used to develop AWQC that can account for the known effects of TMFs on Zn bioavailability.

Supporting Information—The Supporting Information is available on the Wiley Online Library at https://10.1002/etc.5529.

Acknowledgments—We thank C. Mebane and two anonymous reviewers for their thoughtful and helpful reviews of the manuscript. We also thank E. Van Genderen, Ph.D. and C. Cooper, Ph.D. for providing comments on an earlier draft of the manuscript, and the International Zinc Association for funding the present study.

Author Contributions Statement—**David K. DeForest**: Conceptualization; Data curation; Formal analysis; Methodology; Project administration; Writing—original draft; Writing—review & editing. **Adam Ryan**: Conceptualization; Data curation; Formal analysis; Funding acquisition; Methodology; Writing—original draft; Writing—review & editing. **Lucinda M. Tear**: Conceptualization; Formal analysis; Methodology; Validation; Writing—original draft; Writing—review & editing. **Kevin V. Brix**: Conceptualization; Formal analysis; Methodology; Supervision; Writing—original draft; Writing—review & editing.

Data Availability Statement—Data, associated metadata, and calculation tools are available in the Supporting Information or from the corresponding author (davidd@windwardenv.com).

REFERENCES

ASTM International. (2004). *Standard guide for acute toxicity test with the rotifer Brachionus. E1440-91(2004)*. American Society for Testing and Materials.

Besser, J. M., Ivey, C. D., Steevens, J. A., Cleveland, D., Soucek, D., Dickinson, A., Van Genderen, E. J., Ryan, A. C., Schlekat, C. E., Garman, E., Middleton, E., & Santore, R. (2021). Modeling the bioavailability of nickel and zinc to *Ceriodaphnia dubia* and *Neocloeon triangulifer* in toxicity tests with natural waters. *Environmental Toxicology and Chemistry*, *40*, 3049–3062. https://doi.org/10.1002/etc.5178

Brix, K. V., DeForest, D. K., Tear, L. M., Grosell, M., & Adams, W. J. (2017). Use of multiple linear regression models for setting water quality criteria for copper: A complementary approach to the biotic ligand model. *Environmental Science and Technology*, *51*, 5182–5192.

Brix, K. V., DeForest, D. K., Tear, L., Peijnenburg, W., Peters, A., Middleton, E. T., & Erickson, R. (2020). Development of empirical bioavailability models for metals. *Environmental Toxicology and Chemistry*, *39*, 85–100.

Brix, K. V., Tear, L., Santore, R. C., Croteau, K., & DeForest, D. K. (2021). Comparative performance of multiple linear regression and biotic ligand models for estimating the bioavailability of copper in freshwater. *Environmental Toxicology and Chemistry*, *40*, 1649–1661.

Canadian Council of Ministers for the Environment. (2007). *A protocol for the derivation of water quality guidelines for the protection of aquatic life: 2007*. Winnipeg.

Clifford, M., & McGeer, J. C. (2009). Development of a biotic ligand model for the acute toxicity of zinc to *Daphnia pulex* in soft waters. *Aquatic Toxicology*, *91*, 26–32.

Croteau, K., Ryan, A. C., Santore, R., DeForest, D., Schlekat, C., Middleton, E., & Garman, E. (2021). Comparison of multiple linear regression and biotic ligand models to predict the toxicity of nickel to aquatic freshwater organisms. *Environmental Toxicology and Chemistry*, *40*, 2189–2205.

De Schamphelaere, K. A. C., Heijerick, D. G., & Janssen, C. R. (2003). *Development and validation of biotic ligand models for predicting chronic zinc toxicity to fish, daphnids and algae*. Ghent University.

De Schamphelaere, K. A. C., Heijerick, D. G., & Janssen, C. R. (2004). Comparison of the effect of different pH buffering techniques on the toxicity of copper and zinc to *Daphnia magna* and *Pseudokirchneriella subcapitata*. *Ecotoxicology*, *13*, 697–705.

De Schamphelaere, K. A. C., & Janssen, C. R. (2004). Bioavailability and chronic toxicity of zinc to juvenile rainbow trout (*Oncorhynchus mykiss*): Comparison with other fish species and development of a biotic ligand model. *Environmental Science and Technology*, *38*, 6201–6209.

De Schamphelaere, K. A. C., Lofts, S., & Janssen, C. R. (2005). Bioavailability models for predicting acute and chronic toxicity of zinc to algae, daphnids, and fish in natural surface waters. *Environmental Toxicology and Chemistry*, *24*, 1190–1197.

DeForest, D. K., & Van Genderen, E. J. (2012). Application of USEPA guidelines in a bioavailability-based assessment of ambient water quality criteria for zinc in freshwater. *Environmental Toxicology and Chemistry*, *31*, 1264–1272.

DeForest, D. K., Brix, K. V., Tear, L. M., & Adams, W. J. (2018). Multiple linear regression models for predicting chronic aluminum toxicity to freshwater aquatic organisms and developing water quality guidelines. *Environmental Toxicology and Chemistry*, *37*, 80–90.

DeForest, D. K., Brix, K. V., Tear, L. M., Cardwell, A. S., Stubblefield, W. A., Nordheim, E., & Adams, W. J. (2020). Updated multiple linear regression models for predicting chronic aluminum toxicity to freshwater aquatic organisms and developing water quality guidelines. *Environmental Toxicology and Chemistry*, *39*, 1724–1736.

Doherty, J. (2018). *PEST: Model-independent parameter estimation, user manual Part I: PEST, SENSAN and Global Optimisers* (7th ed.). Watermark Numerical Computing.

European Chemicals Agency. 2008. Characterisation of dose [concentration]-response for environment. Guidance on Information Requirements and Chemical Safety Assessment. Helsinki, Finland.

Erickson, R. J., Benoit, D. A., & Mattson, V. R. (1987). *A prototype toxicity factors model for site-specific copper water quality criteria. Revised September 5, 1996. EPA/600/X-87/290*. US Environmental Protection Agency, Environmental Research Laboratory–Duluth.

Garman, E. R., Meyer, J. S., Bergeron, C. M., Blewett, T. A., Clements, W. H., Elias, M. C., Farley, K. J., Gissi, F., & Ryan, A. C. (2020). Validation of bioavailability-based toxicity models for metals. *Environmental Toxicology and Chemistry*, *39*, 101–117.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Heijerick, D. G., De Schamphelaere, K. A. C., & Janssen, C. R. (2002). Predicting acute zinc toxicity for *Daphnia magna* as a function of key water chemistry characteristics: Development and validation of a biotic ligand model. *Environmental Toxicology and Chemistry*, *21*, 1309–1315.

Heijerick, D. G., De Schamphelaere, K. A. C., Van Sprang, P. A., & Janssen, C. R. (2005). Development of a chronic zinc biotic ligand model for *Daphnia magna*. *Ecotoxicology and Environmental Safety*, *62*, 1–10.

Hogstrand, C., Webb, N., & Wood, C. M. (1998). Covariation in regulation of affinity for branchia zinc and calcium uptake in freshwater rainbow trout. *Journal of Experimental Biology*, *201*, 1809–1815.

Naimi, B., Hamm, N. A. S., Groen, T. A., Skidmore, A. K., & Toxopeus, A. G. (2014). Where is positional uncertainty a problem for species distribution modelling? *Ecography*, *37*, 191–203.

Peters, A., Merrington, G., Stauber, J., Golding, L., Batley, G., Gissi, F., Adams, M., Binet, M., McKnight, K., Schlekat, C. E., Garman, E., & Middleton, E. (2021). Empirical bioavailability corrections for nickel in freshwaters for Australia and New Zealand water quality guideline development. *Environmental Toxicology and Chemistry*, *40*, 113–126.

R Core Team. (2021). *R: A language and environment for statistical computing. R version 4.0.4 (2021-02-15)*. R Foundation for Statistical Computing.

Santore, R. C., Di Toro, D. M., Paquin, P. R., Allen, H. E., & Meyer, J. S. (2001). Biotic ligand model of the acute toxicity of metals. 2. Application to acute copper toxicity in freshwater fish and *Daphnia*. *Environmental Toxicology and Chemistry*, *20*, 2397–2402.

Santore, R. C., Mathew, R., Paquin, P. R., & DiToro, D. 2002. Application of the biotic ligand model to predicting zinc toxicity to rainbow trout, fathead minnow, and *Daphnia magna*. *Aquatic Toxicology*, *133*, 271–285.

Soucek, D. J., Dickinson, A., Schlekat, C., Van Genderen, E., & Hammer, E. J. (2020). Acute and chronic toxicity of nickel and zinc to a laboratory cultured mayfly (*Neocloeon triangulifer*) in aqueous but fed exposures. *Environmental Toxicology and Chemistry*, *39*, 1196–1206.

Spry, D. J., & Wood, C. M. (1989). A kinetic method for the measurement of zinc influx *in vivo* in the rainbow trout, and the effects of waterborne calcium on flux rates. *Journal of Experimental Biology*, *142*, 425–446.

Stauber, J., Golding, L., Peters, A., Merrington, G., Adams, M., Binet, M., Batley, G., Gissi, F., McKnight, K., Garman, E., Middleton, E., Gadd, J., & Schlekat, C. (2021). Application of bioavailability models to derive chronic guideline values for nickel in freshwaters of Australia and New Zealand. *Environmental Toxicology and Chemistry*, *40*, 100–112.

US Environmental Protection Agency. (1985). *Guidelines for deriving numerical national water quality criteria for the protection of aquatic organisms and their uses*. Office of Research and Development.

US Environmental Protection Agency. (1987). *Ambient aquatic life water quality criteria for zinc. EPA 440/5-87-003*.

US Environmental Protection Agency. (1996). *1995 updates: Water quality criteria documents for the protection of aquatic life in ambient water. EPA-820-B-96-001*. Office of Water.

US Environmental Protection Agency. (2007). *Aquatic life ambient freshwater quality criteria—Copper. EPA 822-R-07-001*. Office of Water.

US Environmental Protection Agency. (2013). *Aquatic life ambient water quality criteria for ammonia—Freshwater. EPA 822-R-13-001*. Office of Water.

US Environmental Protection Agency. 2015. Toxicity Relationship Analysis Program. Help file for TRAP version 1.3a. Mid-Continent Ecology Division.

US Environmental Protection Agency. (2016). *Aquatic life ambient water quality criteria. Cadmium—2016. EPA-820-R-16-002*. Office of Water.

US Environmental Protection Agency. (2018). *Final aquatic life ambient water quality criteria for aluminum 2018. EPA-822-R-18-001*. Office of Water.

Van Regenmortel, T., Berteloot, O., Janssen, C. R., & De Schamphelaere, K. A. C. (2017). Analyzing the capacity of the *Daphnia magna* and *Pseudokirchneriella subcapitata* bioavailability models to predict chronic zinc toxicity at high pH and low calcium concentrations and formulation of a generalized bioavailability model for *D. magna*. *Environmental Toxicology and Chemistry*, *36*, 2781–2798.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed., p. 495). Springer.

Warne, M. S. J., Batley, G. E., van Dam, R. A., Chapman, J. C., Fox, D. R., Hickey, C. W., & Stauber, J. L. (2018). *Revised method for deriving Australian and New Zealand water quality guideline values for toxicants—Update of 2015 version. Prepared for the revision of the Australian and New Zealand Guidelines for Fresh and Marine Water Quality* (p. 48). Australian and New Zealand Governments and Australian state and territory governments.

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, *1*, 3–14.

