Received: 27 August 2020 | Revised: 12 October 2020 | Accepted: 9 February 2020

Environmental Toxicology

Comparative Performance of Multiple Linear Regression and Biotic Ligand Models for Estimating the Bioavailability of Copper in Freshwater

Kevin V. Brix,[a,b,*] Lucinda Tear,[c] Robert C. Santore,[d] Kelly Croteau,[d] and David K. DeForest[c]

[a]EcoTox, Miami, Florida, USA
[b]University of Miami, Rosenstiel School of Marine and Atmospheric Sciences, Miami, Florida, USA
[c]Windward Environmental, Seattle, Washington, USA
[d]Windward Environmental, Syracuse New York, USA

Abstract: An increasing number of metal bioavailability models are available for use in setting regulations and conducting risk assessments in aquatic systems. Selection of the most appropriate model is dependent on the user's needs but will always benefit from an objective, comparative assessment of the performance of available models. In 2017, an expert workshop developed procedures for assessing metal bioavailability models. The present study applies these procedures to evaluate the performance of biotic ligand models (BLMs) and multiple linear regression (MLR) models for copper. We find that the procedures recommended by the expert workshop generally provide a robust series of metrics for evaluating model performance. However, we recommend some modifications to the analysis of model residuals because the current method is insensitive to relatively large differences in residual patterns when comparing models. We also provide clarification on details of the evaluation procedure which, if not applied correctly, could mischaracterize model performance. We found that acute Cu MLR and BLM performances are quite comparable, though there are differences in performance on a species-specific basis and in the resulting water quality criteria as a function of water chemistry. In contrast, the chronic Cu MLR performed distinctly better than the BLM. Observed differences in performance are due to the smaller effects of hardness and pH on chronic Cu toxicity compared to acute Cu toxicity. These differences are captured in the chronic MLR model but not the chronic BLM, which only adjusts for differences in organism sensitivity. In general, we continue to recommend concurrent development of both modeling approaches because they provide useful comparative insights into the strengths, limitations, and predictive capabilities of each model. *Environ Toxicol Chem* 2021;40:1649–1661. © 2021 The Authors. *Environmental Toxicology and Chemistry* published by Wiley Periodicals LLC on behalf of SETAC.

Keywords: Bioavailability; Metal; Multi-linear regression; Water quality criteria; Biotic ligand model

INTRODUCTION

The influence of toxicity modifying factors (TMFs) such as hardness, pH, and dissolved organic carbon (DOC) on metal bioavailability in aquatic systems has been recognized for over 80 yr (Jones 1939). The US Environmental Protection Agency (1985) developed the first metal bioavailability models to account for the effects of hardness using simple linear regression techniques. Subsequently, more sophisticated mechanistic models were developed that predicted toxicity based on estimations of the free ion concentration of metals (Campbell 1995). These efforts culminated in the development of the biotic ligand model (BLM), which predicts metal accumulation at the gill (or other ligand) of an aquatic organism and relates this accumulation to toxicity (Di Toro et al. 2001).

Biotic ligand models have been developed for a wide range of metals for use by regulators globally (Adams et al. 2020). The BLM has now been widely adopted in Europe for use in regional risk assessments and in setting environmental quality standards. In the United States, the first water quality criteria (WQC) based on the BLM were developed by the US Environmental Protection Agency (2007) for Cu. Since that time, the

This article contains online-only Supplemental Data.

* Address correspondence to KevinBrix@EcoTox.onmicrosoft.com

Published online 16 February 2021 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/etc.5012

Cu BLM has been adopted by many states, but others have been reluctant to do so for a variety of reasons, including lack of clarity on how to implement the model and the perception that the BLM needs too many water quality variables to run the model (Brix et al. 2020).

In response, there has been renewed interest in developing empirical bioavailability models for use in setting WQC and in conducting ecological risk assessments. These statistical models are similar to the earlier hardness-based models except that they frequently take into account multiple TMFs, typically, but not always, using a multiple linear regression (MLR) framework (Brix et al. 2020). This is not a new concept and was first rigorously evaluated by the US Environmental Protection Agency in the mid-1980s on a species-specific basis for Cu using the fathead minnow (*Pimephales promelas*) as a model organism (Erickson et al. 1987). Since then, numerous species-specific models have been developed for a range of metals, as summarized in Brix et al. (2020). Brix et al. (2017) published the first multispecies MLR for Cu in a framework suitable for use in setting environmental quality standards. More recently, the US Environmental Protection Agency (2018) implemented the first MLR-based WQC for Al.

The advantages and disadvantages of mechanistic and empirical models can vary depending on the complexity and interactions of TMFs for a given metal, data availability, as well as intended model use and policy considerations. In 2017, an expert Society of Environmental Toxicology and Chemistry workshop (hereafter, the 2017 Metal Bioavailability Modeling workshop) was held to more critically evaluate approaches and develop quantitative methods for comparing the performance of both mechanistic and empirical bioavailability models (Adams et al. 2020). The objective of the present study was to apply these methods in a comparative analysis of BLM and MLR models for Cu, building on a previous comparison of BLM and MLR models for Cu described in Brix et al. (2017). Copper is one of the most data-rich metals with respect to toxicity data over a range of TMFs, providing an important test of how the standardized procedures for model development, validation, and comparison developed during the 2017 Metal Bioavailability Modeling workshop perform.

METHODS

Toxicity data sets

The starting acute and chronic toxicity data sets for Cu were based on those compiled by the US Environmental Protection Agency (2007) for derivation of the BLM-based WQC. These data sets were subsequently updated by Brix et al. (2017) for development of the original Cu MLR, including more recent studies that meet US Environmental Protection Agency (1985) test acceptability guidelines. A parallel effort to update the Cu toxicity data set was also undertaken as part of efforts to develop a BLM-based WQG for Cu in Canada (Environment and Climate Change Canada 2019). For the present analysis, these 2 data sets were compared and harmonized to a common data set for comparisons of BLM and MLR performance. Several additional studies were identified that were considered of potentially significant value to model development and added to this final harmonized data set (Table 1; Supplemental Data, Table S1).

Model development—BLM

The Cu BLM (US Environmental Protection Agency 2007) for the present analysis was updated to include several adjustments to the binding constants for the biotic ligand binding sites consistent with those used in the development of the BLM-based WQG in Canada (Environment and Climate Change Canada 2019; Supplemental Data, Tables S3 and S4). The initial adjustment was to reduce the pH response of the model. The pH response in the initial Cu BLM calibration was largely determined by an acute fathead minnow data set because there were few data sets available at that time that systematically looked at pH effects on Cu toxicity (Di Toro et al. 2001; Santore et al. 2001; US Environmental Protection Agency 2007). As other data sets that evaluated pH effects on Cu bioavailability became available, it became clear that model performance could be improved by reducing the pH response. With this update, the $Cu(OH)_2$ binding constant was adjusted to flatten the pH response at high pH. Because most of the toxicity data at high pH are generated in high-hardness waters, this change also affected the hardness response. Cation binding constants (for Ca and Mg) were also adjusted to bring the hardness response in the adjusted model closer to that of the original.

Note that the Cu BLM is a model with a single set of binding constants and other parameters that are applied in a consistent manner to both acute and chronic toxicity data. The model is only adjusted for differences in Cu sensitvity among species for a given endpoint and exposure duration. Consequently, any reference to the acute and chronic Cu BLM in the present study refers only to adjustments to the sensitivity parameter rather than 2 truly separate models, as is the case for the acute and chronic MLR models described.

TABLE 1: Summary acute and chronic copper toxicity data sets

Species	Model data set *n* (% total)	Within national range *n* (%)	Validation data set *n*
Acute			
Ceriodaphnia dubia	120 (13%)	71 (59%)	7
Daphnia magna	438 (46%)	294 (67%)	30
Daphnia obtusa	53 (6%)	26 (49%)	3
Daphnia pulex	34 (4%)	13 (38%)	1
Oncorhynchus mykiss	93 (10%)	39 (42%)	4
Pimephales promelas	219 (23%)	131 (60%)	13
Total	957	574 (60%)	57
Chronic			
D. magna	82 (81%)	69 (84%)	8
O. mykiss	19 (19%)	6 (32%)	2
Total	101	74 (74%)	10

Model development—MLR

Development of the MLR models for Cu followed previously described methods (Brix et al. 2017; DeForest et al. 2018; Brix et al. 2020). Three TMFs—hardness, pH, and DOC—were considered, as in previously described MLRs for this metal. Species-specific MLR models with and without TMF interactions were first developed. Species-specific models were developed for those species meeting the required minimum ranges for hardness (100 mg L^{-1}), pH (1.5 units), and DOC (5 mg L^{-1}; Brix et al. 2017). After this initial evaluation, pooled models were developed using previously described methods (Brix et al. 2017, 2020).

Briefly, for both species-specific and pooled models, stepwise linear regression methods were used to evaluate models with only main independent variables (ln[DOC], ln [Hardness], and pH) and models that considered main independent variables and all 2-way interactions. An analysis of covariance (ANCOVA) was conducted to test for differences between species-specific coefficients and the respective mean coefficients of all species in the basic model. The ANCOVA added a "species" term to the model and was run using the data for all species. We specified a deviance contrast matrix for species that tests the value of each species-specific coefficient against the mean of the other species' coefficients for each term. Finally, 2 ANCOVAs, a basic model and a model with interactions, were conducted to select a model to be used in WQC calculations. These models contained an all-species slope for each independent variable and species-specific intercepts.

Best-fitting models were identified using the Akaike information criterion (AIC) and the Bayesian information criterion (BIC; Burnham and Anderson 2004). The model with the smallest AIC or BIC was considered the best-fitting model. Stepwise regressions were run in R using the stepAIC() function from the MASS library (e.g., for the BIC model, stepAIC [model, direction = c("both"), k = 1n(*n*)]).

Model validation

Garman et al. (2020) identified 3 types of validation: 1) autovalidation, 2) independent validation, and 3) cross-species validation. Autovalidation is an evaluation of how well the model describes the toxicity data set from which the model was calibrated or parameterized, whereas independent validation is an evaluation of how well the model developed from 1 toxicity data set predicts toxicity in another data set. Cross-species validation is a specific type of independent validation which evaluates the ability of a model developed for 1 or more species to predict toxicity for a species that is not in the model. The large acute toxicity database for Cu makes it possible to evaluate all 3 types of validation for BLMs and MLR models, though in this particular evaluation we did not consider cross-species validation. Our model validation focused on development of independent data sets for model validation and quantitative evaluation of model performance (i.e., autovalidation).

Independent model validation

Use of data sets independent from those that may have been used to develop a model can be a useful tool for model validation. Previous efforts to develop MLR bioavailability models have generally not included this type of analysis. The acute toxicity data set for Cu is very large and allows for development of independent model validation data sets. These data were removed from the overall toxicity data set prior to developing MLR models. Similarly, for chronic Cu, we were able to remove a small (*n* = 10) independent model validation data set.

The focus of the validation data sets was to understand how MLR and BLM models would predict toxicity in waters that were similar to combinations of TMF concentrations that are known to occur in natural waters of the United States. A data set described in Brix et al. (2020) of 22 087 field-sampled waters from 924 locations around the United States was used to represent water quality conditions (pH, DOC, hardness) to which the models should apply. The mean concentration of each TMF from each location was calculated, and then the locations at which the mean concentration of all TMFs fell between the 2.5 and the 97.5 percentile site mean concentrations (Supplemental Data, Table S2) were retained to create the "US waters" data set (*n* = 809).

To determine which of the water chemistries in each species data set reflect conditions in natural waters, a simple principal component analysis (PCA) of the *Z* scores of log(DOC), log (Hardness), and pH from the final US waters data set was conducted in R (*prcomp*[]; R Development Core Team 2020) to reduce the data set to 2 dimensions. The PCA model was used to predict the axis scores for the log-transformed toxicity data (*predict*[], in R). A scatterplot of the axis 2 versus axis 1 scores of each species toxicity data set was overlain on a scatterplot for the US waters data set to provide a visual assessment of how the scatter of each species' axis scores compared to the scatter for the US waters data set. The toxicity data sets contained a number of studies with TMF combinations that were outside the range of US waters, especially in the upper left quadrant of the plots where DOC concentrations are low. Toxicity data in this region of the plots were excluded from the pool of potential validation data using simple cutoff values from the PCA scores; toxicity studies with PC1 < –2 and PC2 > 2 were excluded. From the possible validation data set for each species, approximately 10% of the toxicity tests were then randomly selected to comprise the validation data set for that species.

Comparative model performance (autovalidation)

The performances of BLM and MLR models were quantitatively characterized and compared using methods described in Brix et al. (2020) and Garman et al. (2020). Performance of the BLM and MLR models refers to the accuracy with which they can predict toxicity over a wide range of bioavailability conditions and can be an important component of overall model validation. For the statistically based MLR models, initial assessments of model performance include standard diagnostic tools for evaluating regressions including adjusted R^2 (squared Pearson correlation coefficient between observed and predicted), predicted R^2, residuals analysis, and variance inflation factors (VIFs; Burnham and Anderson 2004; Zuur et al. 2010; Harrell 2015). Some of these diagnostic tools are also part of the validation steps for comparing BLMs and MLR models.

Garman et al. (2020) provide a general approach for model validation which includes qualitative and quantitative performance evaluations. From this general approach, we conducted the following 4 performance evaluations.

First, we developed 1:1 plots of observed versus predicted effect concentration (EC*x*) values which include diagonal lines denoting 1:1 and a factor of ±2 agreement. Plotting observed versus predicted is preferred because the expected slope of the 1:1 regression line is equal to 1, whereas it is generally <1 when plotted as predicted versus observed (Pineiro et al. 2008). These plots provided an initial assessment of how well the models predict toxicity, characterizing the number of toxicity predictions within a factor of ±2 agreement as well as the degree of scatter and patterns in the toxicity predictions.

Next, we evaluated model residuals of observed versus predicted values as a function of observed toxicity to identify whether there were systematic biases in the model predictions (e.g., underprediction of toxicity at lower observed EC*x* values and overprediction of toxicity at higher EC*x* values). We then evaluated residuals as a function of the TMFs considered in the model (DOC, pH, hardness) to identify the parameter(s) responsible for any observed patterns in residuals.

Next, we developed cumulative probability distributions of observed/predicted EC*x* values and factor-of-agreement plots to expand on information in the 1:1 plots. We included BLM and acute and chronic pooled MLR models along with a null model to provide context on BLM and MLR model performance. The null model was derived by determining the geometric mean of all toxicity values for a species without any bioavailability adjustment. This is the predicted value against which individual observed values were compared (Garman et al. 2020). These plots provide an easy visual comparison of models characterizing the percentage of data that are over- or underpredicted by the models as well as a simple way to evaluate the percentage of toxicity predictions that are within a given factor of agreement.

Finally, these metrics were quantitatively integrated in a comparative model performance evaluation generally following the recommendations in Garman et al. (2020), in which a series of 6 scores were developed: the model (adjusted) R^2, the slope of model residuals versus toxicity, hardness, pH, DOC, and the percentage of data within a factor of 2 ($RF_{x,2.0}$) using the slope rating formula described in Garman et al. (2020). These performance metrics for a model (species-specific or pooled) were summed and divided by the number of metrics evaluated to provide an overall performance index.

In the process of conducting the present analysis, several uncertainties regarding the details of the Garman et al. (2020) procedures were identified that could have significant impacts on the scoring process. Specifically, it was unclear whether the (adjusted) R^2 derived from the 1:1 plots and the residual slope analysis should be based on all data pooled or on a species-specific basis and then averaged across species. We also noted that the equation provided in Garman et al. for scoring residual slopes was not particularly sensitive over the range of slopes observed in the present study. To address this, we also considered an alternative formula

TABLE 2: Summary of species-specific and pooled Cu multiple linear regression models

		Slope					
Species	Intercept	ln(DOC)	ln(Hardness)	pH	*n*	Adjusted R^2	Predicted R^2
Acute							
Ceriodaphnia dubia	–5.427	0.639	0.133	0.985	113	0.71	0.69
Daphnia magna	–4.991	0.746	0.544	0.774	408	0.78	0.78
Daphnia obtusa	–2.063	0.839	0.282	0.551	50	0.82	0.79
Daphnia pulex	0.587	0.763	0.622		33	0.76	0.71
Oncorhynchus mykiss	0.481	0.511	0.758		89	0.58	0.55
Pimephales promelas	–6.654	0.670	0.986	0.962	206	0.77	0.76
Pooled model		0.700	0.579	0.778	899		
C. dubia	–5.890	113	0.64				
D. magna	–5.148	408	0.78				
D. obtusa	–4.942	50	0.64				
D. pulex	–5.373	33	0.69				
O. mykiss	–4.650	89	0.48				
P. promelas	–3.605	206	0.70				
Chronic							
D. magna	0.372	0.851	0.198	0.175	74	0.87	0.86
O. mykiss	–0.064	0.882	0.284	0.284	17	0.64	0.53
Pooled model		0.855	0.221	0.215	91		
D. magna	–0.056	74	0.87				
O. mykiss	0.685				17	0.62	

DOC = dissolved organic carbon.

of 1-|slope| as a scoring tool, with slopes <–1 or >1 constrained to a score of 0.

Model evaluation

Model validation provides information to assess model performance. Following model validation, there may be multiple viable models to consider depending on the needs of the end user. As outlined in Van Genderen et al. (2020), considerations for model selection include representation of the model relative to water chemistry conditions and species to which the model would be applied; level of input required in terms of water chemistry, expertise, operation, and interpretation requirements; model accuracy (as informed by model validation); and ease of model use relative to its purpose.

In the present study, we initially considered using the model evaluation procedure recommended in Van Genderen et al. (2020). However, because our approach involved first developing common data sets for Cu across models, the inputs for all but one of the metrics used in the Van Genderen et al. procedure are the same for the BLM and MLR and provide no discriminatory power. The only metric in this procedure that differs between models is $RF_{x,2.0}$, which is the same metric used in the Garman et al. (2020) procedure for model validation. Although this procedure is likely to be a

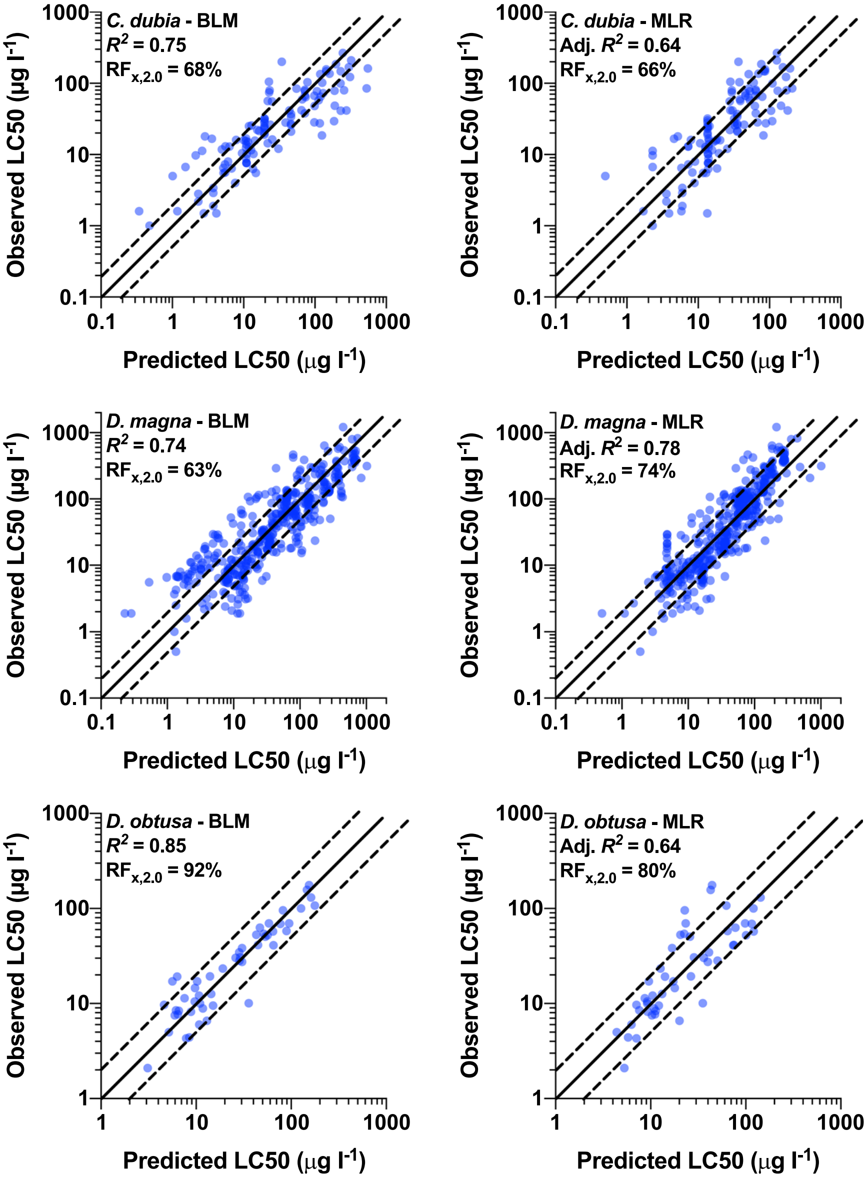

FIGURE 1: Comparative 1:1 plots of the pooled acute Cu multiple linear regression model and the biotic ligand model using autovalidation data sets for *Ceriodaphnia dubia*, *Daphnia magna*, *Daphnia obtusa*, *Daphnia pulex*, *Oncorhynchus mykiss*, and *Pimephales promelas*. Solid line is the line of perfect agreement between predicted and observed median effect concentrations. Dashed lines indicate a factor of ±2. BLM = biotic ligand model; MLR = multiple linear regression; $RF_{x,2.0}$ = percentage of data within a factor of 2 using the slope rating formula; LC50 = median lethal concentration.

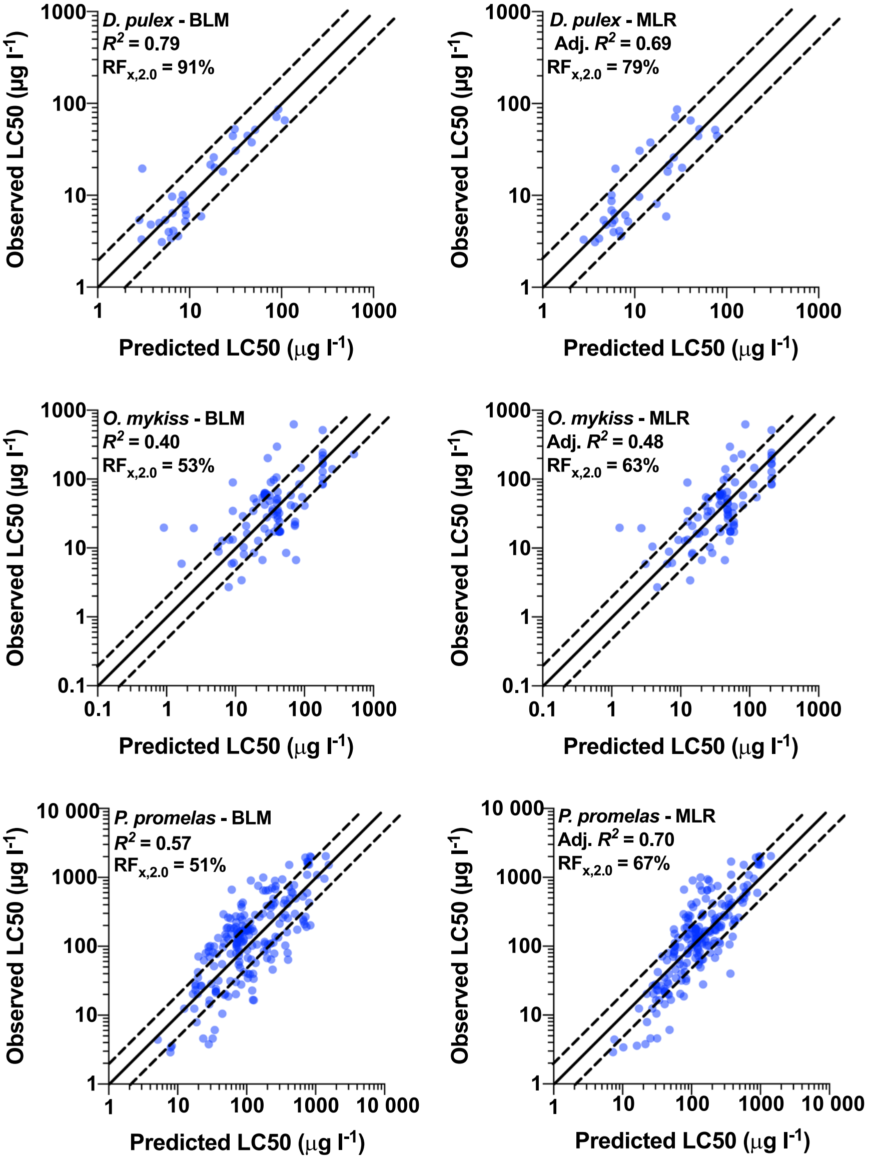

FIGURE 1: (Continued)

useful tool for evaluating models based on different data sets, it does not provide any additional information when data sets have been harmonized between models.

We did, however, consider a complementary analysis to those described in Van Genderen et al. (2020) for comparatively evaluating BLM and MLR models. Specifically, we used the BLM and MLR models described in the present study to derive WQC as a function of the 3 main TMFs (hardness, DOC, and pH). The present analysis used the full species sensitivity distributions (SSDs; Supplemental Data, Tables S4 and S5) to calculate WQC across a range of TMF conditions using standard procedures (US Environmental Protection Agency 1985). We included data from both the auto- and independent validation data sets to ensure that we did not bias the SSDs by withholding data from the analysis. Although this type of analysis does not provide an objective measure or metric of model performance, it does provide important comparative information about how WQC may differ across a range of TMF conditions that model users may find valuable in model selection.

RESULTS

Model development

Six species-specific acute MLR models and 2 chronic MLR models were developed for Cu (Table 2). Four of the acute models were for daphnids (*Ceriodaphnia dubia*, *Daphnia magna*, *Daphnia obtusa*, and *Daphnia pulex*) and the other 2 for fish (*P. promelas* and *Oncorhynchus mykiss*). One chronic daphnid (*D. magna*) and 1 chronic fish (*O. mykiss*) model were also developed. Data from these models were then pooled to develop final acute and chronic pooled models for comparison to the Cu BLM (Table 2). Several of the acute species-specific models (*C. dubia* and *D. obtusa*) performed better when interactions were considered (Supplemental Data, Table S6); but performance for the pooled models with and

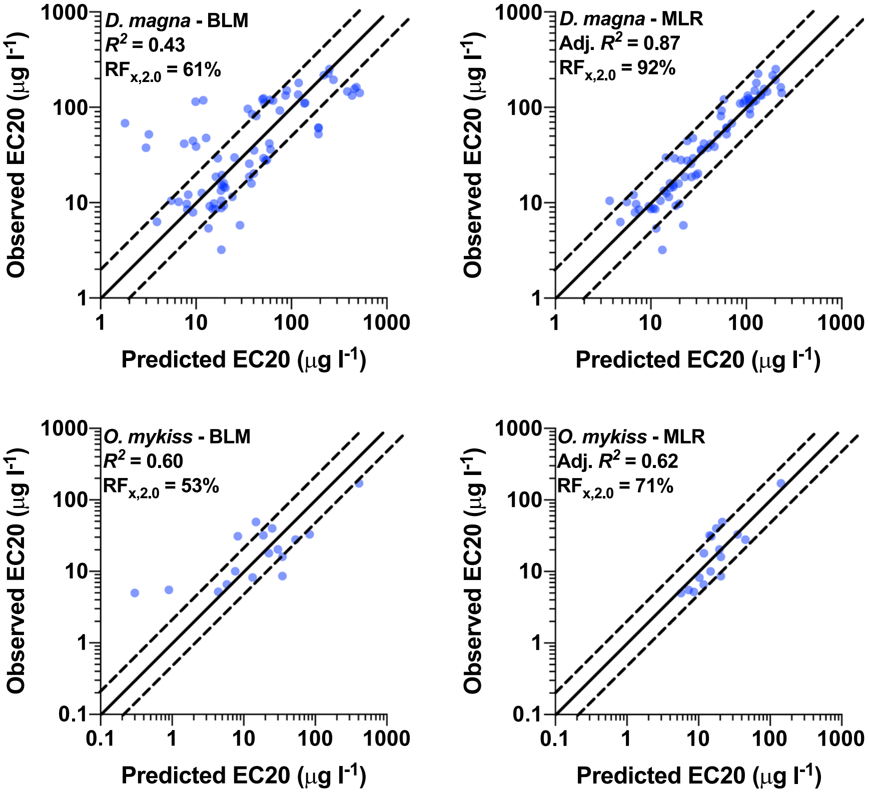

FIGURE 2: Comparative 1:1 plots of the pooled chronic Cu multiple linear regression model and biotic ligand model using autovalidation data sets for *Daphnia magna* and *Oncorhynchus mykiss*. Solid line is the line of perfect agreement between predicted and observed 20% effect concentrations. Dashed lines indicate a factor of ±2. BLM = biotic ligand model; MLR = multiple linear regression; $RF_{x,2.0}$ = percentage of data within a factor of 2 using the slope rating formula; EC20 = 20% effect concentration.

without interactions did not differ, so interaction terms were not retained.

The pooled acute Cu MLR model developed for the present analysis generally performed comparably to the pooled acute Cu MLR model described in Brix et al. (2017). The adjusted R^2 of the pooled acute model applied to the species-specific data sets on average decreased by 0.012 compared to the 2017 Cu MLR model, with *C. dubia* and *D. magna* data sets exhibiting the greatest changes (0.08 and –0.08, respectively). The new acute MLR model for *O. mykiss* did not perform as well as the other species-specific models. The species-specific model had an adjusted R^2 of 0.58, and application of the pooled model to the *O. mykiss* data set resulted in an adjusted R^2 of 0.48 (Table 2). The VIFs, which provide an index that measures how much variance of an estimated regression coefficient is increased by collinearity, were low for hardness, DOC, and pH for the acute species-specific and pooled Cu MLR models, ranging from 1.0 to 2.9 (Supplemental Data, Table S7). Typically, VIFs <3 are considered acceptable (Zuur et al. 2010).

As with the MLR, the updated Cu BLM performed similarly, on average, to the US Environmental Protection Agency (2007) Cu BLM used in our previous comparison of Cu BLM and acute MLR models (Brix et al. 2017; Figure 1). The R^2 of the BLM applied to the species-specific acute data sets on average decreased by 0.002 compared to the analysis in Brix et al. (2017). Also similar to the MLR, application of the BLM to the acute *O. mykiss* data resulted in a relatively low R^2 of 0.40.

The chronic pooled Cu MLR model developed for the present analysis included the addition of *O. mykiss*, whereas the previous chronic MLR developed in Brix et al.

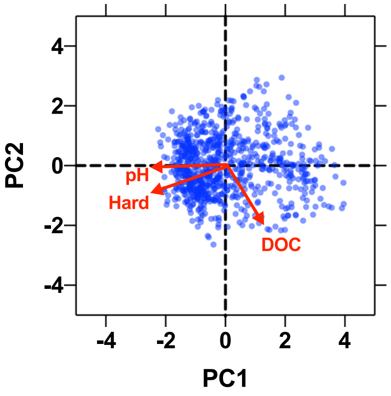

FIGURE 3: Biplot of axis scores for principal components 1 and 2 of principal component analysis of *Z* scores of log(DOC), log(Hardness), and pH from the natural waters data set. Numbers represent stations, and red arrows show the direction of increasing concentrations of associated toxicity modifying factors. DOC = dissolved organic carbon; PC = principal component.

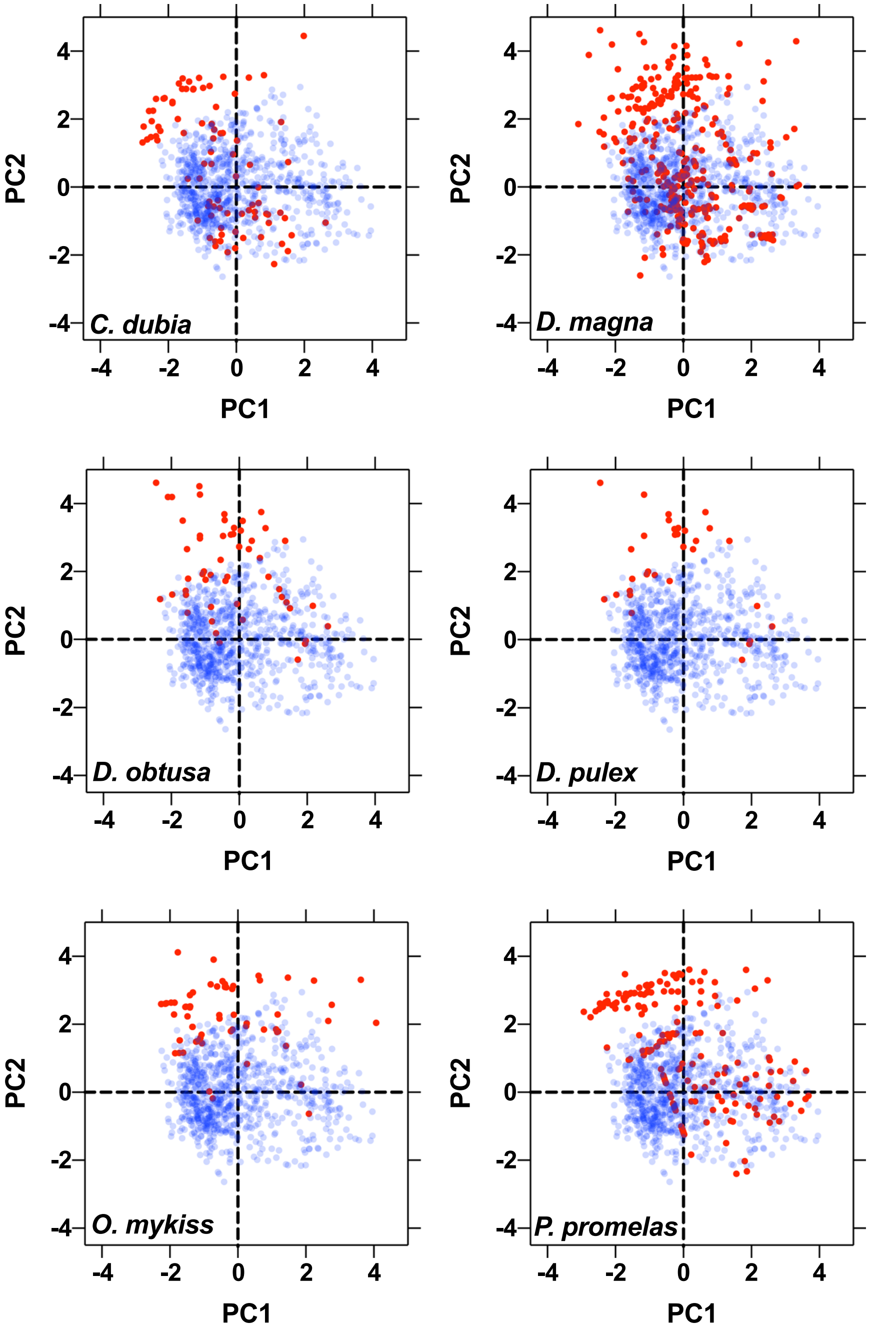

FIGURE 4: Comparison of predicted axis scores of acute toxicity data (red circles) to scores of principal components 1 and 2 of the natural waters data set (light blue circles). PC = principal component; *C. dubia* = *Ceriodaphnia dubia*; *D. magna* = *Daphnia magna*; *D. obtusa* = *Daphnia obtuse*; *D. pulex* = *Daphnia pulex*; *O. mykiss* = *Oncorhynchus mykiss*; *P. promelas* = *Pimephales promelas*.

(2017) was based solely on data for *D. magna*. The updated species-specific chronic Cu MLR model for *D. magna* had an adjusted R^2 of 0.87, which is the same as the adjusted R^2 in Brix et al. (2017). Application of the pooled chronic Cu MLR model to the chronic *D. magna* data set had an adjusted R^2 of 0.87 (Figure 2). Application of the BLM to the chronic *D. magna* data set resulted in a lower R^2 of 0.43. For the chronic *O. mykiss* data set, application of the pooled MLR model resulted in an adjusted R^2 of 0.62, and the BLM resulted in an R^2 of 0.60 (Figure 2). The VIFs for hardness, DOC, and pH were low for the chronic species-specific and pooled Cu MLR models, ranging from 1.0 to 2.0 (Supplemental Data, Table S7), indicating low collinearity between TMFs.

Comparison of TMFs for toxicity data and natural waters

Evaluation of toxicity data sets for model development and validation data involved comparing TMFs for available toxicity data relative to the distribution of TMFs in natural waters in the United States. The first 2 PCA axes of the US natural water data set explained 92% of the variance (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; Figure 3). The toxicity data sets used to develop the models contained a number of studies with TMF combinations that were outside the range of US waters, especially in the upper left quadrant of the plots where DOC concentrations are low. Only 38 to 67% (60% on average) of the data for any given species data set was representative of US waters (upper left and upper right quadrants; Figure 4 and Table 1). Toxicity data in this region of the plots were excluded from the pool of potential validation data using simple cutoff values from the PCA scores; toxicity studies with PC1 <–2 and PC2 >2 were excluded. From the possible validation data set for each species, approximately 10% of the toxicity tests were then randomly selected to comprise the validation data set for that species (Table 1). The specific data used in the validation data sets can be found in Supplemental Data, Table S8.

Comparative model performance

Of the 6 acute species-specific data sets that were evaluated in the autovalidation data set, the BLM performed better for 3 species (*C. dubia*, *D. obtusa*, and *D. pulex*), and the MLR performed better for the other 3 species (*D. magna*, *O. mykiss*, and *P. promelas*; Figure 1). The adjusted R^2 for the pooled MLR ranged from 0.48 to 0.78 for the species-specific data sets with $RF_{x,2.0}$, ranging from 63 to 80%. The BLM R^2 ranged from 0.40 to 0.85 for the species-specific data sets, with 51 to 92% of predictions within a factor of 2 of observed (Figures 1 and 5).

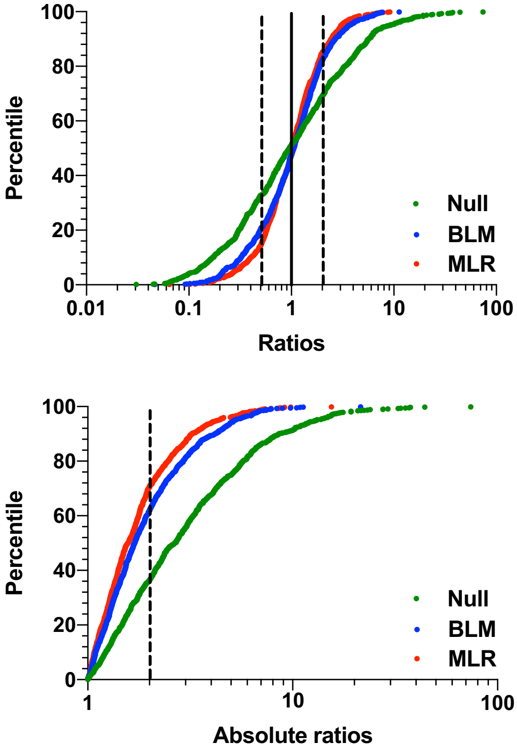

FIGURE 5: Cumulative density functions of predicted to observed ratios for acute Cu null, biotic ligand, and multiple linear regression models. Data plotted using actual and absolute values for all *Ceriodaphnia dubia*, *Daphnia magna*, *Daphnia obtusa*, *Daphnia pulex*, *Oncorhynchus mykiss*, and *Pimephales promelas* data. Ratios calculated as predicted/observed. Absolute ratios calculated as $10^{|\log(\text{predicted/observed})|}$. Dashed lines indicate a factor of ±2. BLM = biotic ligand model; MLR = multiple linear regression.

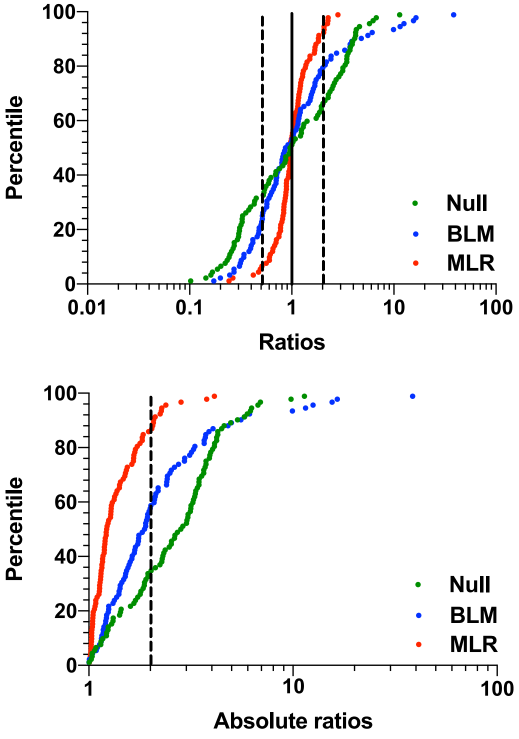

FIGURE 6: Cumulative density functions of predicted to observed ratios for chronic Cu null, biotic ligand, and multiple linear regression models. Data plotted using actual and absolute values for all *Daphnia magna* and *Oncorhynchus mykiss* data. Dashed lines indicate a factor of ±2. BLM = biotic ligand model; MLR = multiple linear regression.

The chronic pooled Cu MLR performed better than the BLM for the *D. magna* data set, with an adjusted R^2 of 0.87 and $RF_{x,2.0}$ of 92% compared to an R^2 of 0.43 and $RF_{x,2.0}$ of 61% for the BLM. The results were mixed for the *O. mykiss* data set, with the BLM having a comparable R^2 (0.60) but lower $RF_{x,2.0}$ (53%) than the MLR (adjusted R^2 = 0.62, $RF_{x,2.0}$ = 71%; Figures 2 and 6).

For the acute Cu BLM and pooled Cu MLR model, results from the analysis of residual slopes varied depending on the method used. In general, both BLM and MLR performances were scored lower using the mean of slopes for individual species compared to slopes for all data pooled, but in some cases, substantial differences were found between the scoring methods (Supplemental Data, Table S9). For example, the residual slope analysis for predicted median lethal concentrations (LC50s) versus hardness indicated that the MLR and BLM are comparable using the mean slope from individual species (differences in mean slope ~0.01) but that the MLR performs substantially better than the BLM based on the slope analysis of pooled data (1.0 and 0.73, respectively, based on the 1-|slope| scoring method; Supplemental Data, Table S9). Further, the Garman et al. (2020) method for slope scoring consistently resulted in higher scores and often indicated more comparable performance between the BLM and MLR models compared to the 1-|slope| scoring method.

Based on the scoring analysis using the mean of individual species slopes, the BLM performed better than the acute MLR when analyzing residuals versus observed toxicity, whereas MLR performed better when analyzing residuals versus DOC (Supplemental Data, Table S9 and Figures S1 and S3). The 2 models were comparable for hardness and pH based on mean scores, though there were substantial differences when evaluating individual species (Supplemental Data, Figures S2 and S4). The overall scores generated by taking the average score of all metrics were identical (<0.01 difference) when based on the mean of individual species slopes, indicating that overall the BLM and acute MLR model performances are quite comparable (Supplemental Data, Table S9).

The scoring analysis for the chronic Cu data set again indicated that the 1-|slope| scoring method had higher discriminatory power than the Garman et al. (2020) method. The MLR scored better than the BLM across all of the metrics (Supplemental Data, Table S10). The residual slope score for pH was particularly low for the BLM (Supplemental Data, Table S10 and Figure S6). Overall, using the 1-|slope| scoring method and the mean of individual species score, the BLM had a score of 0.55 compared to 0.87 for the MLR.

Model validation using independent data sets

Analysis of the acute Cu independent validation data set yielded somewhat ambiguous results. The relatively small sample sizes for individual species precluded use of the coefficient of determination (R^2) statistic. Using the pooled acute validation data set, $RF_{x,2.0}$ was 74% for the BLM and 81% for the pooled MLR model (Figure 7). For the chronic independent validation data set, the MLR model performed better than the BLM, with an $RF_{x,2.0}$ of 80% compared to 60% for the BLM; but the data set was small ($n = 10$; Figure 8).

Comparative model evaluations

As discussed, we did not focus our model evaluation on the methods described in Van Genderen et al. (2020) because the methodology was not designed to compare models based on the same toxicity data set. Instead, we conducted a comparative model evaluation by deriving WQC across a range of TMF conditions.

We evaluated both acute and chronic WQC. The BLM is the same for both acute and chronic WQC, with only the sensitivity adjusted, whereas for the MLR there are separate models. Interestingly, the BLM and MLR show close correspondence for hardness and DOC based on the chronic MLR and for pH based on the acute MLR (Figure 9). Compared to the acute MLR, the BLM has a shallower response for hardness and a steeper response for DOC, whereas compared to the chronic MLR, the BLM has a steeper response to pH.

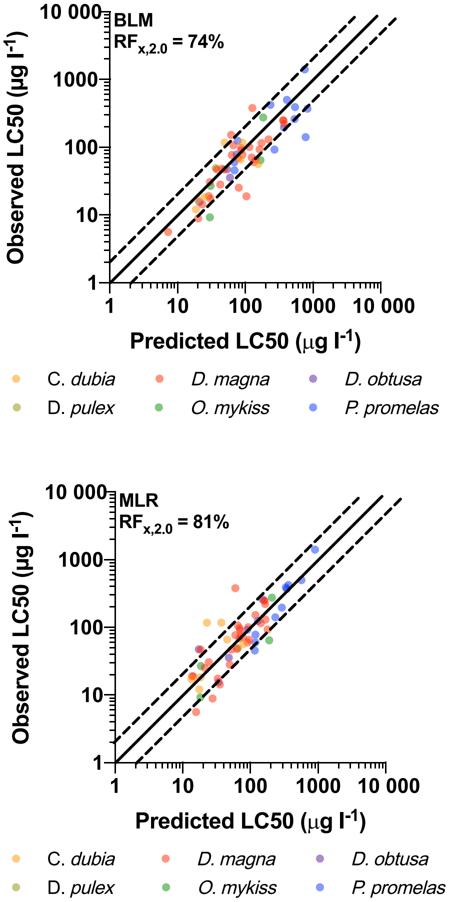

FIGURE 7: Comparative 1:1 plots of the pooled acute Cu multiple linear regression model and biotic ligand model using independent data sets for *Ceriodaphnia dubia*, *Daphnia magna*, *Daphnia obtusa*, *Daphnia pulex*, *Oncorhynchus mykiss*, and *Pimephales promelas*. Solid line is line of perfect agreement between observed and predicted 20% effect concentrations. Dashed lines indicate a factor of ±2. BLM = biotic ligand model; $RF_{x,2.0}$ = percentage of data within a factor of 2 using the slope rating formula; LC50 = median lethal concentration; MLR = multiple linear regression.

DISCUSSION

An increasing number of models are being developed to predict metal toxicity as a function of multiple TMFs. These models span the spectrum from empirical to mechanistic with a variety of gradations in between (Adams et al. 2020; Brix et al. 2020). Selecting the most appropriate model for a given application requires consideration of a number of factors including data requirements and availability, the intended model use, various practical and policy considerations, and overall model performance. The objective of the present study was to evaluate this last factor—model performance—by applying procedures developed as part of an expert workshop to a representative metal that has relatively well-developed BLM and MLR models. Most of the metrics recommended from this workshop were already available but have been used inconsistently in characterizing the performance of metal bioavailability models across studies. The workshop also developed 2 novel model scoring schemes that integrate

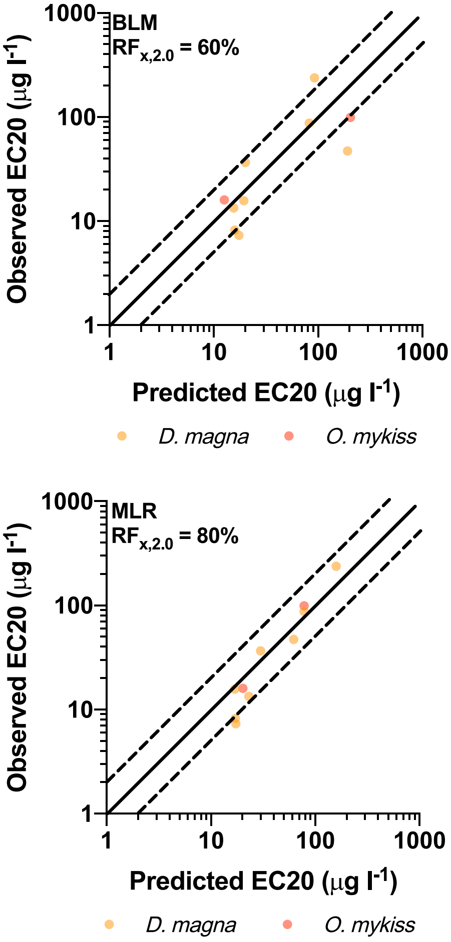

FIGURE 8: Comparative 1:1 plots of the pooled chronic Cu multiple linear regression model and the biotic ligand model using independent data sets for *Daphnia magna* and *Oncorhynchus mykiss*. Solid line is the line of perfect agreement between observed and predicted EC20s. Dashed lines indicate a factor of ±2. BLM = biotic ligand model; $RF_{x,2.0}$ = percentage of data within a factor of 2 using the slope rating formula; EC20 = 20% effect concentration.

multiple metrics into a single score (Garman et al. 2020; Van Genderen et al. 2020). Our analysis provides information on both the comparative performance of BLMs and MLR models for Cu and a critical evaluation of the model comparison procedures developed during the 2017 Metal Bioavailability Modeling workshop.

Our analysis indicates that the acute Cu BLM and MLR model perform quite similarly based on summary metrics such as the (adjusted) R^2, $RF_{x,2.0}$, and the total score method of Garman et al. (2020) when considering mean responses across species. However, substantial differences in model performance exist when individual species and metrics are considered. For example, the BLM performs substantially better than the MLR in predicting toxicity for the *D. pulex* and *D. obtusa* data sets, whereas the MLR performs better for the *P. promelas* data set (Figure 1). As another example, the residuals plotted as a function of DOC have a mix of relatively small positive and negative slopes for the MLR but are consistently larger and all negative for the BLM, suggesting a consistent bias in the BLM model for DOC (Supplemental Data, Figure S3). The source of this bias may be related to assumptions in the BLM regarding the fractions of humic and fulvic acids that comprise DOC but was not investigated further in our analysis.

Our analysis of the Cu BLM and MLR models for predicting chronic toxicity indicates that the MLR generally performs better than the BLM. However, the Cu BLM is not optimized for the chronic toxicity observations (i.e., the same BLM parameters are used for acute and chronic toxicity data), whereas the chronic Cu MLR is based explicitly on chronic Cu toxicity data, so it is not surprising that it performs better than the BLM. It should also be noted that comparison of the 2 models for predicting chronic toxicity is somewhat less robust than for the acute Cu data set because data were limited to 2 species (*D. magna* and *O. mykiss*). Keeping this caveat in mind, our analysis of available chronic Cu toxicity data suggests that the slopes for hardness and pH are substantially lower in the pooled chronic MLR model compared to the pooled acute MLR model (Table 2). These apparent differences are captured by the MLR model but not the BLM, which assumes that TMFs behave in the same manner for acute and chronic toxicity and that only the sensitivity parameter is adjusted.

The evaluation methodologies developed during the 2017 Metal Bioavailability Modeling workshop were applied in our analysis. We found the model scoring scheme described in Garman et al. (2020) to be useful in providing an integrated measure of model performance. However, we also identified several steps in the process where additional detail and method development may be warranted. First, Garman et al. are unclear on whether the various scoring metrics should be conducted on a pooled data set or on data sets for individual species and then integrated by a summary statistic such as the mean. We evaluated both approaches and concluded that application of the scoring procedure to the pooled data set is generally not appropriate. Use of (adjusted) R^2 on the pooled data set is not appropriate because it increases variance (i.e., the total sum of squares) in the independent variable (e.g., LC50 or 20% effect concentration) because of differences in species sensitivity. This will lead to an increase in the (adjusted) R^2 that does not reflect an improved ability to account for effects of TMFs on metal toxicity. Analysis of residuals on the pooled data set is also inappropriate because it can obscure patterns in the residuals for individual species. For example, plotting residuals versus hardness for individual species using the Cu MLR results in slopes ranging from –0.38 to 0.49. In contrast, if the data are pooled across species, these meaningful differences in slopes in the residuals are obscured by the mixture of positive and negative slopes and differences in species sensitivity resulting in a slope of –0.0003 (Supplemental Data, Table S9). Consequently, analysis of the pooled data set obscures the fact that the model may be normalizing data for a particular species in a biased manner as a function of one or more TMFs.

We also considered an alternative (1-|slope|) to the scoring statistic for residuals recommended in Garman et al. (2020).

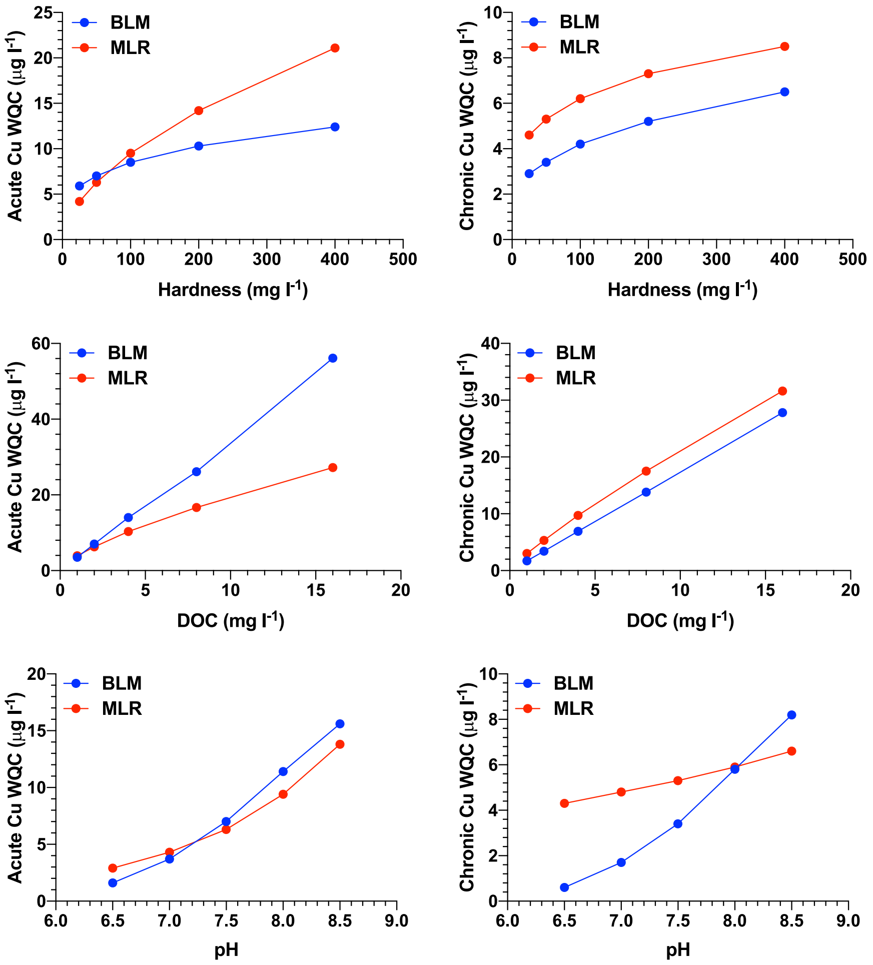

FIGURE 9: Comparison of acute and chronic water quality criteria (WQC) for Cu based on biotic ligand and multiple linear regression models as a function of hardness, dissolved organic carbon (DOC), and pH. Hardness-dependent WQC calculated at DOC = 2 mg L^{-1}, pH = 7.5. DOC-dependent WQC calculated at hardness = 50 mg L^{-1}, pH = 7.5. pH-dependent WQC calculated at hardness = 50 mg L^{-1}, DOC = 2 mg L^{-1}. BLM = biotic ligand model; MLR = multiple linear regression.

There is no "best" statistic to use in this type of analysis because it depends on how sensitively the user wants larger slopes (positive or negative) in the residuals to be penalized. However, we observed that the statistic recommended in Garman et al., when applied to log-transformed data, as was used in deriving the models, was not very sensitive to relatively large differences in residual slopes. For example, a slope of 1 in the residuals, which is larger than any slope we observed in our data, would receive a score of 0.5 (from a possible range of 0–1) using Garman et al., whereas it would receive a score of 0 using the 1-|slope| statistic.

Our evaluation of the scoring system described in Van Genderen et al. (2020) indicates that this scoring system is not appropriate when comparing models that have been developed using the same toxicity data set. When this is the case, all of the scoring metrics are the same except for $RF_{x,2.0}$, which becomes the sole basis for model selection. It is apparent that the authors developed this scoring system to evaluate models that were independently developed using different toxicity data sets. When this is the case, the Van Genderen et al. methodology may be very useful; but it has yet to be robustly tested.

CONCLUSION

The present study provided a comparative evaluation of BLM and MLR models for Cu based on methods developed in the 2017 Metals Bioavailability Modeling workshop. Overall, we found the methodology developed in Garman et al. (2020) to be useful in discriminating differences in model performance, although some modifications and clarifications to the methodology are recommended. In contrast, the methodology described in Van Genderen et al. (2020) is not appropriate when models are developed using a common toxicity data set but may be appropriate in scenarios where models are developed independently using different toxicity data sets.

Our analysis led us to conclude that the acute Cu BLM and MLR model performances are quite comparable overall, similar to our previous conclusion using few metrics (Brix et al. 2017), but that the chronic Cu MLR performs better than the chronic Cu BLM, which is the same model as the acute Cu BLM but adjusted for sensitivity. However, despite similarities in model performance for acute Cu toxicity, there are meaningful differences in the WQC derived for both acute and chronic Cu toxicity.

In the last several years, there has been an increase in the number of environmental regulatory agencies that are considering use of MLR models as an alternative to the BLM. The relative simplicity of the model form (though we would argue that MLR model development can be as complicated as the BLM) and its generally comparable or, in some cases, better performance make it an attractive alternative to the BLM. Although this may be the case, we continue to advocate for the concurrent development of both BLM and MLR models when possible. In our view, MLR models should be treated as complementary simplifying regulatory tools of scientifically more robust mechanistic bioavailability models such as the BLM.

Supplemental Data—The Supplemental Data are available on the Wiley Online Library at https://doi.org/10.1002/etc.5012.

Acknowledgment—The authors acknowledge A. Edgington for assistance in compiling toxicity databases. The International Copper Association provided financial support for this research.

Data Availability Statement—Data, associated metadata, and calculation tools are available from the corresponding author (KevinBrix@EcoTox.onmicrosoft.com).

REFERENCES

Adams WJ, Blust R, Dwyer R, Mount DM, Nordheim E, Rodriguez PH, Spry DJ. 2020. Bioavailability assessment of metals in freshwater environments: A historical review. *Environ Toxicol Chem* 39:48–59.

Brix KV, DeForest DK, Tear L, Grosell M, Adams WJ. 2017. Use of multiple linear regression models for setting water quality criteria for copper: A complementary approach to the biotic ligand model. *Environ Sci Technol* 51:5182–5192.

Brix KV, DeForest DK, Tear L, Peijnenburg WJGM, Peters A, Middleton ET, Erickson RJ. 2020. Development of empirical bioavailability models for metals. *Environ Toxicol Chem* 39:85–100.

Burnham KP, Anderson DR. 2004. Multimodel inference: Understanding AIC and BIC in model selection. *Sociol Methods Res* 33:261–303.

Campbell PGC. 1995. Interactions between trace metals and aquatic organisms: A critique of the free-ion activity model. In Tessier A, Turner DR, eds, *Metal Speciation and Bioavailability Aquatic Systems*. John Wiley and Sons, New York, NY, USA, pp. 45–102.

DeForest DK, Brix KV, Tear LM, Adams WJ. 2018. Multiple linear regression (MLR) models for predicting chronic aluminum toxicity to freshwater aquatic organisms and developing water quality guidelines. *Environ Toxicol Chem* 37:80–90.

Di Toro DM, Allen HE, Bergman HL, Meyer JS, Paquin PR, Santore RC. 2001. A biotic ligand model of the acute toxicity of metals. I. Technical basis. *Environ Toxicol Chem* 20:2383–2396.

Environment and Climate Change Canada. 2019. Draft Federal Environmental Quality Guidelines—Copper. Gatineau, QC, Canada.

Erickson RJ, Benoit DA, Mattson VR. 1987. A prototype toxicity factors model for site-specific water quality criteria. US Environmental Protection Agency, Duluth, MN.

Garman ER, Meyer JS, Bergeron CM, Blewett TA, Clements WH, Elias MC, Farley KJ, Gissi F, Ryan AC. 2020. Validation of bioavailability-based toxicity models for metals. *Environ Toxicol Chem* 39:101–117.

Harrell FE Jr. 2015. *Regression Modeling Stategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer, Cham, Switzerland.

Jones JRE. 1939. Antagonism between salts of the heavy and alkaline-earth metals in their toxic action on the tadpole of the toad, *Bufo bufo bufo* (L.). *J Exp Biol* 16:313–333.

Pineiro G, Perelman S, Guerschman JP, Paruelo JM. 2008. How to evaluate models: Observed vs. predicted or predicted vs. observed? *Ecol Modell* 216:316–322.

R Development Core Team. 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Santore RC, Di Toro DM, Paquin PR, Allen HE, Meyer JS. 2001. Biotic ligand model of the acute toicity of metals. 2. Application to acute copper toxicity in freshwater fish and *Daphnia*. *Environ Toxicol Chem* 20:2397–2402.

US Environmental Protection Agency. 1985. Guidelines for deriving numerical national water quality criteria for the protection of aquatic organisms and their uses. Duluth, MN.

US Environmental Protection Agency. 2007. Aquatic life ambient freshwater quality criteria—Copper. Washington, DC.

US Environmental Protection Agency. 2018. Final aquatic life ambient water quality criteria for aluminum 2018. Washington, DC.

Van Genderen EJ, Stauber JL, Delos CG, Eignor D, Gensemer RW, McGeer JC, Merrington G, Whitehouse P. 2020. Best practices for derivation and application of thresholds for metals using bioavailability-based approaches. *Environ Toxicol Chem* 39:118–130.

Zuur AF, Ieno EN, Elphick CS. 2010. A protocol for data exploration to avoid common statistical problems. *Methods Ecol Evol* 1:3–14.