Environmental Toxicology and Chemistry—Volume 40, Number 8—pp. 2189–2205, 2021
Received: 25 August 2020 | Revised: 2 October 2020 | Accepted: 1 April 2021

Environmental Toxicology

Comparison of Multiple Linear Regression and Biotic Ligand Models to Predict the Toxicity of Nickel to Aquatic Freshwater Organisms

Kelly Croteau,[a,*] Adam C. Ryan,[b] Robert Santore,[a] David DeForest,[c] Christian Schlekat,[d] Elizabeth Middleton,[d] and Emily Garman[d]

[a]Windward Environmental, Syracuse, New York, USA
[b]International Zinc Association, Durham, North Carolina, USA
[c]Windward Environmental, Seattle, Washington, USA
[d]NiPERA, Durham, North Carolina, USA

Abstract: Toxicity-modifying factors can be modeled either empirically with linear regression models or mechanistically, such as with the biotic ligand model (BLM). The primary factors affecting the toxicity of nickel to aquatic organisms are hardness, dissolved organic carbon (DOC), and pH. Interactions between these terms were also considered. The present study develops multiple linear regressions (MLRs) with stepwise regression for 5 organisms in acute exposures, 4 organisms in chronic exposures, and pooled models for acute, chronic, and all data and compares the performance of the Pooled All MLR model to the performance of the BLM. Independent validation data were used for evaluating model performance, which for pooled models included data for organisms and endpoints not present in the calibration data set. Hardness and DOC were most often selected as the explanatory variables in the MLR models. An attempt was also made at evaluating the uncertainty of the predictions for each model; predictions that showed the most error tended to show the highest levels of uncertainty as well. The performances of the 2 models were largely equal, with differences becoming more apparent when looking at the performance within subsets of the data. *Environ Toxicol Chem* 2021;40:2189–2205. © 2021 SETAC

Keywords: Water quality guidelines; Biotic ligand model; Metal toxicity; Multiple linear regression

INTRODUCTION

It has long been recognized that metal toxicity to aquatic organisms can be affected by a number of water quality parameters; in particular, nickel water quality guidelines in the United States and Canada have been based on hardness equations for several decades (US Environmental Protection Agency 1996). The hardness equation is an example of a simple empirical regression that characterizes how nickel toxicity is reduced in waters with elevated hardness. More recently, the biotic ligand model (BLM) has been proposed as a way to explain, at a more mechanistic level, how certain water quality constituents affect metal toxicity. The hardness effect, for example, has been shown to result from competition between nickel and hardness cations (calcium and magnesium) on gill surfaces (Meyer et al. 1999), and this information was used to parameterize a competitive reaction in an early version of a nickel BLM (Water Environment Research Federation et al. 2003). Although the hardness equation and the BLM are attempts to model the same thing, they represent 2 very different approaches: the empirical approach is a simple equation that is built directly from observations of toxicity over ranges of hardness, whereas the more complex mechanistic approach of the BLM is based on simulating the chemical interactions underlying the way in which hardness cations affect the quantity of nickel that causes toxicity. Both approaches have their merits, and both can be used in the development of water quality guidelines. For example, as noted, the hardness equation has been in use for several decades for nickel and other metals, and the BLM was used in the development of the US Environmental Protection Agency (USEPA) water quality criteria for copper (US Environmental Protection Agency 2007).

It is important to note that other bioavailability normalization approaches exist for nickel, including the one described by Nys et al. (2016). This approach serves as the basis for the bioavailability-based environmental quality standard

This article includes online-only Supplemental Data.
* Address correspondence to kellyc@windwardenv.com

Published online 13 April 2021 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/etc.5063

for nickel under the European Water Framework Directive (European Commission 2011). In accordance with European guidelines, the Nys et al. (2016) approach focuses primarily on the use of chronic data in species-specific models. Recently, a nickel guideline was developed in Australia by Stauber et al. (2021) that was based on trophic level–specific multiple linear regressions (MLRs) developed by Peters et al. (2021) with ecotoxicity data from species that would be relevant to Australia and New Zealand. The Stauber et al. (2021) analysis also considered the Peters et al. (2021) pooled MLR, the Nys et al. (2016) bioavailability model, and the Peters et al. (2018) bioavailability model. The approach that has been adopted in the United States and Canada combines data from broad taxonomic groups into a single "pooled" model for either acute data, chronic data, or all data (Brix et al. 2017; DeForest et al. 2017; Environment and Climate Change Canada 2019). Given the scope of the present study, it is justifiable that the MLR developed within this document should be compared to a BLM that was developed following similar assumptions.

As new information becomes available, guidelines need to be updated to accurately represent the best science and be protective of aquatic life. Given the availability of both empirical and mechanistic methods for incorporating bioavailability effects, it is necessary to perform a comprehensive comparison of the merits and limitations of each approach to assist regulators in guideline development.

In the present study, the empirical and mechanistic approaches for assessing water quality factors affecting nickel toxicity are explored and evaluated. Empirical models are not limited to hardness; rather, they take the form of MLRs to evaluate which toxicity-modifying factors (TMFs) need to be incorporated in the analysis. The approach for developing MLRs follows that of Brix et al. (2017, 2020) and DeForest et al. (2018). The BLM for nickel has most recently been described by Santore et al. (2021, in this issue) and uses a set of biotic ligand binding constants that have been calibrated to fit toxicity data, making it a mechanistic model in concept but with some empirical calibration done to achieve the final fit. The objectives of our study are to review freshwater toxicity data for fish, invertebrates, plants, and algae to compile a database for quantifying nickel effects in acute and chronic exposures; develop acute and chronic nickel MLR models in a manner consistent with Brix et al. (2017) and DeForest et al. (2018); compare and evaluate the nickel BLM and nickel MLR approaches to the same data sets; and quantify MLR- and BLM-based outcomes in a manner consistent with objective comparative approaches proposed by Garman et al. (2020) and van Genderen et al. (2020).

METHODS

Toxicity database

A database containing nickel effect concentrations (EC*x*) and all relevant associated exposure conditions was compiled to include data from all identified primary sources (i.e., published studies with aquatic nickel toxicity data). The AQUIRE database and Google Scholar were used to identify relevant literature to include in the database. Data were screened using a quality scoring process comparable to that used to identify data suitable for inclusion in water quality guidelines (US Environmental Protection Agency 1985; Canadian Council of Ministers of the Environment 2007). The criteria for selecting data for guideline development typically included numbers of replicates, frequency of nickel measurements in the test, reporting of control performance, reporting of the statistical methods used, and particular criteria regarding the length of the test and biological endpoints deemed acceptable for the intended application. Because the purpose of this analysis was to develop and test models—not to develop a water quality guideline—the standards for acceptable data were not as strict, and they primarily focused on having enough information about the test conditions (e.g., documented water quality) to make MLR and BLM predictions, as well as enough information about the test organisms to classify to at least the genus level. Concentrations lethal to 50% of a population based on a single exposure concentration measurement (for calibration data) or nominal concentrations (for validation data), endpoints that would not be chosen for guideline development (e.g., chronic survival endpoints), and studies that did not report replication and statistical methods were all considered acceptable for the purposes of model development and validation because they still provided valuable information about the trends and relative influences of TMFs. Although more than 300 articles were considered for inclusion in the database, data from studies considered to be of insufficient quality were not included. To facilitate development of MLR models capable of characterizing the effects of various TMFs on EC*x* values, exposure conditions associated with each EC*x* were described in as much detail as possible. In addition, because MLR model performance was to be compared to BLM performance, data or estimates for all BLM inputs were necessary. In many cases, the detailed water chemistry data necessary for BLM calculations were not available. For example, many studies reported pH, alkalinity, and hardness but not concentrations of individual constituents such as dissolved calcium, magnesium, or sodium. In such cases, it was necessary to use the available water chemistry data for each study to estimate detailed water chemistry. After screening for data quality and sufficient reported water quality information to support the application of BLM and MLR procedures, data from 81 studies were considered for inclusion in the database (Supplemental Data 1).

The procedure for estimating detailed chemistry values in the database is described in the BLM user's guide (Windward Environmental 2017). The reported chemistry was used whenever possible, and the estimation procedure was used with ion ratios assigned on a case-by-case basis with information from the source water, as is recommended in the BLM user guide (Windward Environmental 2017). Estimations of missing chemical parameters included cases wherein major ions, alkalinity, or dissolved organic carbon (DOC) were missing. Studies that did not include enough information to allow for this estimation procedure or that did not report pH were not considered except in a few cases where pH was estimated. A study would normally be rejected if the pH was not reported because there was typically no good basis for estimating the pH in a test water, but there were 5 instances in the final data set for which the authors were confident enough to estimate the pH. Alsop and Wood (2011) required the pH to be estimated for 3 of their exposures because they reported the pH of a control synthetic water and then the modifications to that water (i.e., salt addition); the pH of the modified water was assumed to be the same as that of the control water. Ferreira et al. (2010), Keller and Zam (1991), Kim et al. (2017), and Pavlaki et al. (2011) did not report pH measurements for any of their exposures; instead, they used a standard water recipe—ASTM International hard water for Ferreira et al. (2010) and Pavlaki et al. (2011) and USEPA soft or moderately hard water for Keller and Zam (1991) and Kim et al. (2017)—for which a typical range of pH values was given in the methods for making the water; the average pH from that range was used as the estimate. Importantly, all of these were acute exposures where the pH was not expected to change during the test.

If ion concentrations and alkalinity measurements were not available, values for these parameters were estimated from other reported sources for the same water type. For example, the ion concentrations for studies that reported using synthetic water recipes, such as USEPA hard water or ASTM International hard water, were estimated from the synthetic water recipe (US Environmental Protection Agency 1994; ASTM International 1996). If only one of the major ions (i.e., calcium, magnesium, sodium, potassium, sulfate, or chlorine) was then missing, the concentration of the missing ion was estimated based on charge balance. If full chemistry was not yet achieved, then ion ratios were assigned based on the source water, with median ion ratios from North America assigned if no other source-specific values could be determined (see Supplemental Data 2 for a summary of the ion ratios used). Hardness was sometimes calculated from the sum of calcium and magnesium concentrations if only individual ion concentrations were reported. In the final data set, hardness needed to be estimated in 8 of the studies (67 data points), not including studies that reported calcium and magnesium but not hardness. If alkalinity could not be estimated from other sources, alkalinity (as dissolved inorganic carbon) was estimated to be in equilibrium with the atmosphere, with an atmospheric carbon dioxide pressure of $10^{-3.2}$ atm.

Reported values for major ions and organic carbon that were given in total rather than dissolved terms were assumed to be equivalent to the dissolved values (total = dissolved). The exceptions were nickel concentrations which were converted with a total-to-dissolved conversion factor of 0.998 (US Environmental Protection Agency 1996).

If the DOC concentration was not reported for a toxicity test, then an attempt was made to assign a reasonable value based on the type of exposure media used. Synthetic test media without dissolved organic matter (DOM) added were assumed to have 0.3 mg DOC/L (Santore et al. 2002) regardless of whether the test was fed or not. We made this assumption because it was not always clear if the test was fed, the DOC concentration in a fed test would not be so very much higher that it would greatly influence the DOC trend to use this assumption, and using this lower value would be more conservative. Also assumed was a composition of 0.01% humic acid in synthetic waters because all DOC present would have been contributed by the organisms or food sources and would more likely resemble fulvic acid rather than terrestrially derived humic acid (US Environmental Protection Agency 2007). The percentage of humic acid for natural water sources was estimated as 10%, except when the authors had reported adding humic acid specifically (e.g., adding DOM as Aldrich Humic Acid), in which case the humic acid percentage was calculated based on those reported carbon additions, as appropriate. If the source water for a given study was characterized in other published literature, the average DOC concentrations from those sources were used as the DOC estimate. Supplemental Data 3 provides a complete list of standard DOC values that were used for estimation. In the final data set, DOC was estimated in 28 studies (295 data points), only 7 of which (34 data points) involved estimating DOC for natural waters.

There have been some questions in the past about using results from exposures wherein the pH was buffered with nonnatural buffers (e.g., 3-[*N*-morpholino]propanesulfonic acid [MOPS]) because such buffers have been shown to change the slope of the effect of pH on toxicity (Kozlova et al. 2009; Esbaugh et al. 2013). This subject was investigated as part of the present study, but details were omitted for brevity; specifics are available in Supplemental Data 4. Briefly, the pH slope was significantly different between daphnid exposures that buffered with MOPS and those that did not; however, the effect on the final pooled models was minor, especially when the final pooled model did not include pH as an explanatory term. Furthermore, the loss of data from excluding any exposure that used a synthetic buffer was undesirable because it would prevent the development of a model for algae and it is impractical to control pH in an algae study without the use of a buffer. This analysis therefore continued to use data buffered with MOPS, including 74 data points used for calibration, 33 used for primary validation, and 6 used for secondary validation (Parametrix 2005; De Schamphelaere et al. 2006; Deleebeeck et al. 2007, 2008a, 2008b, 2009; Kozlova et al. 2009; AECOM Technical Services 2011). These data points included those for *Ceriodaphnia dubia*, *Daphnia magna*, *Daphnia pulex*, *Oncorhynchus mykiss*, and *Pseudokirchneriella subcapitata*.

Selection of data for model development and validation

Of the 1588 observations in the compiled toxicity and chemistry database, 910 either had fewer than 2 observations available to calculate a species mean value, did not have chemistry that varied enough within a group, or had a more desirable endpoint available (e.g., concentrations effective on 50% of a population [EC50s] were preferentially used). A single observation would obviously give no information about model performance using the methods described in the present study, and any variation in groups that did not have varying chemistry would be purely due to random biological variability, not bioavailability relationships. The order of preference for endpoints (from most desirable to least) was EC50, EC20, EC10, maximum allowable toxicant concentration, lowest-observed-effect concentration, and no-observed-effect concentration. The EC50 was most preferable because values generally vary less from the fit of the response curve, making them better to use in predictions when evaluating models. Of the remaining observations, 242 were selected for calibrating the MLR model, and 436 were selected for validation (81 for primary validation and 355 for secondary validation). The final data set for MLR development and comparison to the BLM (Supplemental Data 1) included 678 toxicity observations from 59 articles: 300 from acute tests and 378 from chronic tests. The acute data comprised data from 27 species from 21 genera: 5 fish and 22 invertebrate species. The chronic data comprised data from 28 species from 21 genera: 1 amphibian, 2 fish, 17 invertebrate, 3 plant, and 5 algae species.

The purpose of defining primary and secondary validation data sets was to use as much of the available data as possible to independently evaluate model performance. Primary validation data were chosen to represent the species, exposure durations, and endpoints that were considered during model calibration so that calibrated MLRs could be applied to the data directly, thereby testing the consistency of both the bioavailability relationships and the sensitivity parameter. Restricting the comparison to a validation data set that included only data that matched species, life stage, and endpoint would have been preferable; but the primary validation data that met these conditions resulted in too few data for a meaningful comparison. Secondary validation data for the species-specific models were chosen to represent the species and test types (i.e., acute or chronic) that were considered during calibration, but life stages, exposure durations, and endpoints may have differed from those in the calibration data set. Consequently, the intercepts of the calibrated MLRs were adjusted to an optimal value when applied to the secondary validation data, but all coefficients from the calibrated model were used. The intent of the secondary validation was to use available data to further evaluate the bioavailability effects (i.e., effects of TMFs) described by the calibrated models beyond the comparisons afforded by primary validation data availability. The pooled models were tested with the same secondary validation data as the species-specific models, as well as with data from species that did not have enough data for a species-specific model. Noncalibration species were only used when validating the pooled models because these models are meant to be more generally applicable, and they span a larger range of EC, values so there is less risk of the additional data artificially inflating the R^2 statistics. These data were once again grouped by species, life stage, endpoint, and duration and were used to evaluate the pooled model of the appropriate test type. The ranges of chemical parameters that would be relevant to MLR development are summarized in Figure 1 for all calibration and validation data (similar summaries for each of the calibration groups described in Table 1 are available in Supplemental Data 5). Figure 1 also shows how the ranges of chemistry compare to the 5 to 95% range of those parameters in US surface waters (Hirsch and De Cicco 2015).

Calibration data were selected so that a wide range of chemistry conditions would be considered for each organism, life stage, test duration, and endpoint. In addition, data from tests in synthetic water were preferred for inclusion in the calibration data set, whereas data from tests in natural waters were preferred for inclusion in the primary validation data set. This preference was because tests in synthetic waters tend to be more abundant, and they are able to systematically change single parameters. Natural waters tend to have a more random assortment of chemistry, making them good for validation but harder to examine for potential relationships for calibration purposes. Some test waters were classified as "altered natural," including waters that had been derived from natural sources but were processed in some way prior to testing (e.g., adding salts to a natural water or treating a natural water to remove some components, including dechlorinated tap water). These altered natural waters were assigned to the calibration data set if the calibration data set did not represent a wide enough range of chemistry conditions or to the primary validation data set if the number of observations in the primary validation data set was relatively small.

Selection criteria for what constituted a wide enough range of chemistry for the calibration data set were consistent with those described in Brix et al. (2020): the range of pH in the data had to span at least 1.5 standard units, the range of DOC had to span at least 5 mg C/L, and the range of hardness had to span at least 100 mg/L as calcium carbonate. The DOC range was typically the most difficult criterion to meet because relatively few tests conducted in synthetic water contained added organic matter, and data from natural water tests were initially reserved for the validation data set. If the DOC criterion could not be met using tests conducted in synthetic water, then 2 of the observations from the natural waters were chosen with a random number generator and reassigned to the calibration data set. The selection criteria were met for each species, with the exception of acute *Daphnia pulicaria* data and chronic *O. mykiss* data. Chronic *O. mykiss* data were included even though the DOC spanned a range of only 4.57 mg C/L. The *D. pulicaria* data were included even though the hardness spanned a range of only 95 mg/L as calcium carbonate and only natural waters were available, so two-thirds of the data (11 of 16) were randomly chosen for use in the calibration data set, leaving approximately one-third (5 of 16) for validation. The ranges of water chemistry in the acute and chronic data that were used for model development and evaluation are listed in Table 2, as are the ranges of the same parameters in US surface waters. The full range of chemistry present in actual US surface waters goes well beyond the

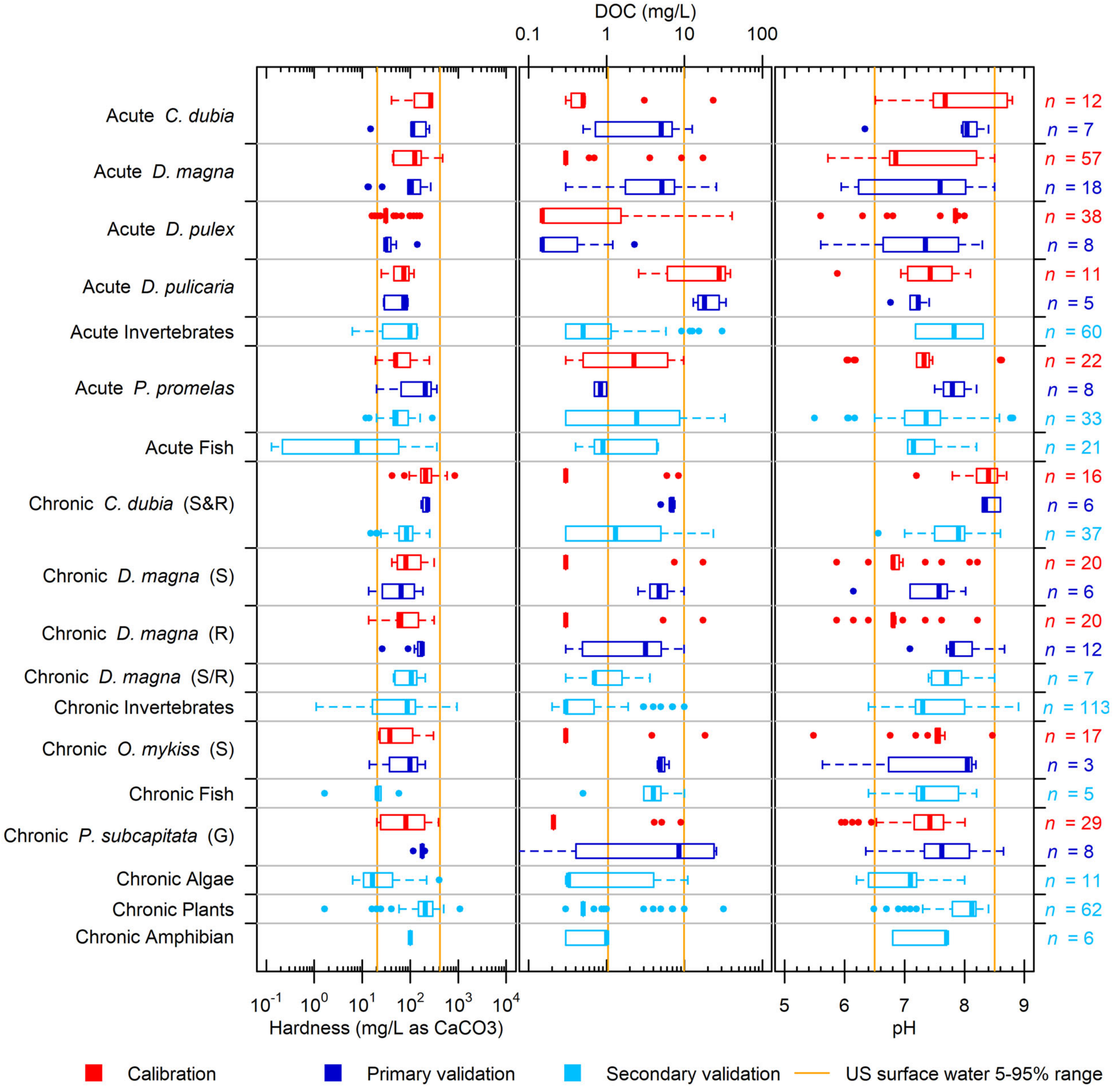

FIGURE 1: Summary of dissolved organic carbon, hardness, and pH data used in the Pooled All model calibration and validation data sets. Vertical lines represent the 5th to 95th percentile range for parameters within US waters. DOC = dissolved organic carbon; S = survival; R = reproduction; G = growth.

bounds of what is typically tested in a laboratory setting. However, at least 90% of the chemistry measurements (the 5–95% range) fell within the range represented in the toxicity data. The exception was acute validation data, for which the maximum hardness value used in the toxicity data was only 360 mg $CaCO_3$/L, compared to the US surface water 95th percentile of 415 mg $CaCO_3$/L.

MLR calibrations and predictions

The MLR models were fit using methods consistent with those described in Brix et al. (2017) and DeForest et al. (2018). Briefly, the step() function in R (R Development Core Team 2020) was used to determine the best-fit set of parameters. The step() function starts with a null model (i.e. toxicity is not affected by modifying factors) and adds parameters up to, at most, a "full-scope" model equation with all of the allowed parameters. The program decides whether to keep the current model equation, add parameters, or eliminate parameters one at a time by comparing their Akaike information criterion (AIC) or Bayesian information criterion (BIC). The AIC and BIC are both based on information theory and are measures of the relative quality of a model; these criteria differ only in their value of *k* in the equation used to calculate them ([–2 × log-likelihood] + [*k* × npar], where npar is the number of parameters in the model; Sakamoto et al. 1986). Whereas the AIC uses

TABLE 1: Number of toxicity observations in each group for the calibration and validation data sets

		n		
Test type	Group	Calibration	Primary validation	Secondary validation
Acute	*Ceriodaphnia dubia* 48-h LC50	12	7	—
Acute	*Daphnia magna* 48-h LC50	57	18	—
Acute	*Daphnia pulex* 48-h LC50	38	8	—
Acute	*Daphnia pulicaria* 48-h LC50	11	5	—
Acute	Other invertebrates	—	—	60
Acute	*Pimephales promelas* 96-h LC50 (adult)	22	8	33
Acute	Other fish	—	—	21
Acute	Pooled acute	140	46	114
Chronic	*C. dubia* 7-d survival and reproduction IC25	16	6	37
Chronic	*D. magna* 21-d reproduction MATC	20	6	7[a]
Chronic	*D. magna* 21-d survival MATC	20	12	7[a]
Chronic	Other invertebrates	—	—	113
Chronic	*O. mykiss* 17- to 26-d LC50	17	3	—
Chronic	Other fish	—	—	5
Chronic	*Pseudokirchneriella subcapitata* 72-h growth EC50	29	8	—
Chronic	Other algae	—	—	11
Chronic	Other plants	—	—	62
Chronic	Other amphibians	—	—	6
Chronic	Pooled Chronic	102	35	241
Acute + chronic	Pooled All	242	81	355

[a] Because both are chronic *D. magna* groups, the secondary validation data set can be applied to either. These are the same 7 observations.
LC50 = concentration lethal to 50% of a population; IC25 = concentration required for 25% inhibition of an effect; MATC = maximum allowable toxicant concentration; EC50 = concentration effective on 50% of a population.

$k = 2$, the BIC uses $k = \ln(n)$, where n is the number of observations; thus, the BIC penalizes the number of parameters more harshly when the data set is large (AIC and BIC values improve as they approach negative infinity).

Four scenarios were run based on all 4 combinations of 2 different conditions: whether interaction terms were considered for inclusion and whether the best fit was determined based on the AIC or the BIC. In 2 of the scenarios, the initial model did not consider interaction terms, meaning that the full-scope equation took the following form:

$$\ln(\text{NiEC}x) = a_D \times \ln(\text{DOC}) + a_H \times \ln(\text{hardness}) + a_p \times \text{pH} + b \quad (1)$$

Three additional terms were included in the full-scope equation for the 2 scenarios that considered the interaction terms, so the equation took the following form:

$$\begin{aligned}\ln(\text{NiEC}x) = {} & a_D \times \ln(\text{DOC}) + a_H \times \ln(\text{hardness}) + a_p \times \text{pH} \\ & + a_{DH} \times [\ln(\text{DOC}) \times \ln(\text{hardness})] \\ & + a_{Dp} \times [\ln(\text{DOC}) \times \text{pH}] \\ & + a_{Hp} \times [\ln(\text{hardness}) \times \text{pH}] + b \quad (2)\end{aligned}$$

This procedure was performed for each group in Table 1 and for the 3 pooled groups in an analysis of covariance

TABLE 2: Ranges of water chemistry in US surface waters and the acute and chronic data sets used for model development and evaluation

	US surface water		Acute data range		Chronic data range	
Parameter	Range	5–95% range	Calibration	Validation	Calibration	Validation
pH	2.9–10.3	6.5–8.5	5.6–8.8	4.1–8.8	5.5–8.7	3.5–8.9
DOC (mg C/L)	0.1–284	1.0–9.9	0.2–41	0.2–34	0.2–18	<0.1–32
Hardness (mg $CaCO_3$/L)	0.1–4440	21–415	16–477	0.13–360	14–848	1.1–1100
Alkalinity (mg $CaCO_3$/L)	0.1–1970	17–240	0.1–387	0.4–376	<0.1–197	<0.1–229
Temperature (°C)	0–37.2	1.8–29	4–25	10–28	15–26	15–29
Calcium (mg/L)	<0.1–404	5.4–105	0.8–181	<0.1–96	3.0–237	<0.1–392
Magnesium (mg/L)	<0.1–866	1.3–38	0.2–111	<0.1–41	1.1–94	0.2–114
Sodium (mg/L)	0.1–7460	1.6–160	0.5–322	0.6–190	1.8–108	1.1–305
Potassium (mg/L)	<0.1–262	0.6–9.2	<0.1–31	<0.1–12	0.4–157	<0.1–117
Sulfate (mg/L)	0.1–2150	2.7–308	3.2–890	1.3–242	4.0–790	0.5–393
Chloride (mg/L)	<0.1–15 200	0.7–177	0.2–519	<0.1–156	0.6–318	0.4–550

DOC = dissolved organic carbon.

(ANCOVA), with the group included as a categorical parameter (i.e., the intercept would be different in each group, but the slopes would be the same for all groups). The predictions for the calibration and primary validation data sets used the intercepts directly from the models developed during calibration. The intercepts for the secondary validation data set were subsequently calculated for each group. The intercept back-calculated for each observation was calculated as ln(Ni EC*x*) minus the sum of variables times their coefficients; this value then was averaged across all points in a group. So, for a calibrated model containing only DOC and hardness terms, the intercept would be calculated as follows:

$$b_{group} = \left[\sum_{i=1}^{n_{group}} (\ln(\text{NiEC}x)_i - (a_D \times \ln(\text{DOC})_i + a_H \times \ln(\text{hardness})_i)) \right] \Big/ n_{group} \quad (3)$$

The adjusted R^2 and predicted R^2 were calculated for each of the model's calibration data sets. The adjusted R^2 penalizes the normal R^2 based on the number of parameters in the model (k) and the number of calibration data points (n) so that the improvement in model fit is balanced with increases in model complexity.

$$\text{Adjusted } R^2 = 1 - \frac{SSE/(n-k-1)}{SSE/(n-1)} = 1 - \frac{\sum_{i=1}^{n}[(x_i - \hat{x}_i)^2]/(n-k-1)}{\sum_{i=1}^{n}[(x_i - \overline{x})^2]/(n-1)} \quad (4)$$

where SSE is the sum of squared error $\sum_{i=1}^{n}[(x_i - \hat{x}_i)^2]$ and SST is total sum of squares $\sum_{i=1}^{n}[(x_i - \overline{x})^2]$. The predicted R^2 is a way to tell if a model has been overparameterized (i.e., is specific to the data on which it was based instead of being generally applicable; Allen 1974; Tarpey 2000; Hopper 2014). Essentially, it is a form of cross-validation that evaluates how well a particular model may be expected to predict responses for new observations. The process for calculating predicted R^2 involves systematically removing each observation from the model, refitting the coefficients, predicting the response for the observation that was removed, and calculating residuals for an R^2 calculation. The predicted R^2 can range from negative infinity to 1, although like the adjusted R^2, it cannot be higher than the R^2. A predicted R^2 that is much lower than the R^2 or adjusted R^2 is indicative of an overfitted model. Although the predicted R^2 is most useful as a cross-validation technique when an independent validation data set is not available, it is useful in the present context because many of the primary validation data sets are very small. An R^2 was also calculated from the predictions of the primary and secondary validation data sets, separately and together.

The 95% confidence intervals were calculated for each prediction. Calculating the confidence intervals for the calibration/primary validation groups required very little effort because the predict() function in R could return confidence intervals with its predictions for these groups. However, for the secondary validation data, for which new intercepts were calculated, the predictions were calculated manually with the fitted slopes and calculated intercepts. As a result, the confidence intervals also had to be calculated manually. The confidence interval for secondary validation data sets was calculated as follows

$$\hat{y}_i \pm t_{\frac{\alpha}{2},df} * SE(\hat{y}_i) = \hat{y}_i \pm t_{\frac{\alpha}{2},N-n_{par}-1} * \sqrt{\frac{s^2_{b_{group}}}{n_{group}} + \Sigma_{j=1}^{n_{par}} (SE(a_j)_* |x_{i,j} - \overline{x}_{calib,j}|)^2}$$

where $t_{\frac{\alpha}{2},df}$ is the quantile from the t distribution at $\frac{\alpha}{2}$%, degrees of freedom equal to the $s^2_{b_{group}}$ is the variance of the new intercept calculated from Equation 3, $SE(a_j)$ is the standard error of the slope of parameter j, $x_{i,j}$ is the value of parameter j, and $\overline{x}_{calib,j}$ is the average of that parameter from the model's calibration data set.

The final model selection for each calibration group was based largely on the various R^2 statistics described in the present study. If the improvement in the model fit was only slight or if the different R^2 values were leading to inconsistent conclusions (e.g., the primary validation R^2 being higher but the secondary validation R^2 being lower), then some consideration was given to which model had fewer explanatory variables. This consideration was similar in concept to looking at the adjusted R^2—basically, that more complex models should be penalized so that the gains in model performance would be significant enough to justify the addition of more variables.

BLM predictions

The BLM used is the most up-to-date model for predicting the toxicity of nickel to freshwater organisms (Santore et al. 2021, in this issue). The nickel BLM was not calibrated directly to the data composing the calibration data set for the MLRs. Rather, it was previously calibrated to a subset of the MLR calibration data set. Specifically, the binding constants were set with the series of exposures varying the relevant parameter (e.g., calcium for the biotic ligand [BL]-calcium constant, magnesium for the BL-magnesium constant) in the studies listed in the "Calibration data sets" column of Table 3. The particular data points used are indicated by a column in Supplemental Data 1, when possible. The binding constants used in the BLM applied in the present study are provided in Table 3. The BLM binding constants were originally set with a quantitative structure–activity relationship approach; then, each parameter was systematically adjusted to minimize the error in predictions for toxicity tests, which were conducted to evaluate the effects of varying individual water chemistry characteristics (e.g., pH, DOC, calcium, magnesium) on nickel toxicity (Santore et al. 2021, in this issue). Significant differences between fish and invertebrate or acute and chronic data sets were not seen when calibrating binding constants, and the benefits of having a larger calibration data set outweigh any potential

TABLE 3: Binding constants for pooled fish and invertebrate data in the biotic ligand model analysis

Reaction	Log *K*	Contributor to toxicity	Calibration data sets
BL × Ni = BL + Ni^{2+}	4.00	Yes	*Pimephales promelas*, accumulation data (Meyer et al. 1999)
BL × NiOH = BL + $NiOH^{\mp}$	4.357	Yes	
BL × $NiHCO_3$ = BL + $NiHCO_3^+$	5.71 (*C. dubia* only)	Yes (*C. dubia* only)	*Ceriodaphnia dubia* (Schubauer-Berigan et al. 1993; Parametrix 2005)[a]
BL × Ca = BL + Ca^{2+}	4.25	No	*Oncorhynchus mykiss* (Deleebeeck et al. 2007) *Daphnia magna* (Deleebeeck et al. 2008a) *Daphnia pulex* (Kozlova et al. 2009) *P. promelas* (Meyer et al. 1999)
BL × Mg = BL + Mg^{2+}	3.60	No	*O. mykiss* (Deleebeeck et al. 2007) *D. magna* (Deleebeeck et al. 2008a) *D. pulex* (Kozlova et al. 2009)
BL × Na = BL + Na^+	1.00	No	—
BL × H = BL + H^+	4.70	No	—

[a] Made use of unpublished data from Parametrix (2005) study.
BL = biotic ligand.

improvements to the fit of the model from splitting the data set. In addition, sufficient algae data were not available at the time of calibration to inform the calibrations; but plant and algae data that later became available were included in the validation data set, and good fits were seen in those data. A binding constant for nickel bicarbonate binding to the BL is also included in Table 3, although this reaction was not used in the MLR to BLM comparisons. It is included to show a possible alternative model that can provide a better fit for some species, such as *C. dubia*, that seem to show a stronger response to pH than do most other organisms.

The BLM predictions were calculated by first applying the BLM to the MLR calibration data set in speciation mode. Outputs from applying the BLM in speciation mode were processed to calculate the critical accumulation for each observation, which was the concentration of nickel bound to the BL and directly associated with the observed toxic effect. The geometric mean of the critical accumulations in a group was used as the group's final critical accumulation, unless the geometric mean was outside of the interquartile range (IQR; i.e., less than the 25th percentile or greater than the 75th percentile), in which case the median was used. This approach was used because if the geometric mean was outside the IQR, then it was likely to be heavily influenced by an outlier observation. The median (i.e., 50th percentile) was less sensitive to outliers and a better estimate of the center of the data in those cases. In most cases, the geometric mean was used. Once the final critical accumulation was determined for each group in Table 3, the BLM was run in toxicity mode using the test chemistry and the group-specific critical accumulation to determine a predicted effect concentration.

Confidence intervals were determined for the model predictions by first calculating the confidence intervals of the critical accumulation value for each group. The 95% confidence intervals of the critical accumulation geometric mean values were calculated as the 2.5 and 97.5% quantiles from a nonparametric bootstrapping of the geometric mean critical accumulation. For each group, n_{group} critical accumulation values were resampled, with replacement, 1000 times. The geometric mean of the critical accumulation values was calculated for each of these iterations, and the 2.5 and 97.5% quantiles were calculated. This yielded upper and lower confidence limits on the geometric mean critical accumulation that, when run through the BLM in toxicity mode, gave the upper and lower confidence limits of the predicted EC*x* values.

Comparing MLR and BLM performance

For the purposes of comparing the 2 models, the Pooled All MLR model (calibrated with all data pooled) was compared to the single calibrated BLM. The Pooled All MLR model was chosen for comparison because it was more consistent with the BLM application, which assumes a consistent set of binding constants for all organisms for both acute and chronic tests. It should be noted that although the BLM binding constants were calibrated to data from fish and invertebrates only, this is only because sufficient data were not available at the time to include a plant or alga in the calibration data set, but the model has since been validated against plant and alga data with acceptable results ($R^2 = 0.745$). Other comparisons could be performed with the organism-specific MLR models or the Pooled Acute and Pooled Chronic MLR models, but these should be compared to BLMs with similarly specific binding constants.

Fit statistics were calculated for the calibration data set to evaluate the MLR models, but MLR and BLM performances were only compared with the primary and secondary validation data sets (Figure 2), to achieve a fairer comparison. Although some of the data used to calibrate the BLM were included in the MLR calibration data set, the data sets were not identical, which could have biased a comparison of the models based on the calibration data set in favor of one of the models. Furthermore, an empirical model like the MLR model will usually fit its calibration data set better than a mechanistic model like the BLM; a mechanistic model must maintain a predefined structure based on the concepts it is built with, so any changes to

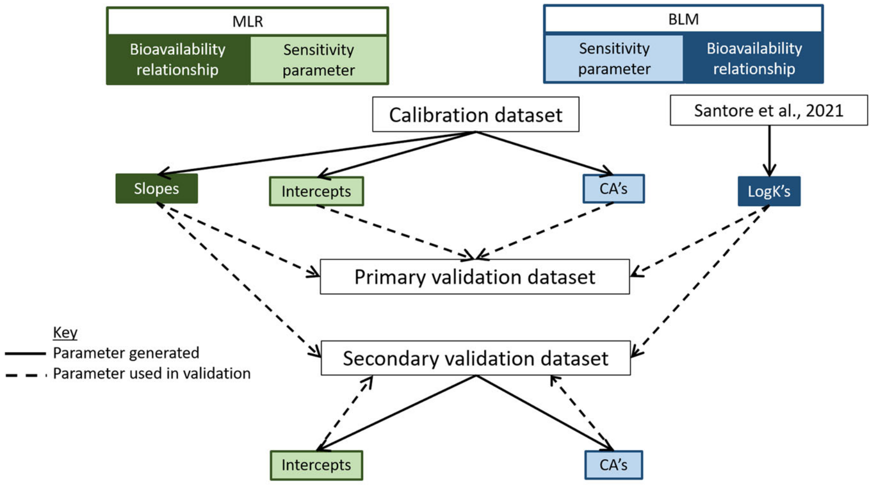

FIGURE 2: Schematic representation of multiple linear regression (MLR) model development and evaluation of MLR model and biotic ligand model performance with calibration, primary validation, and secondary validation data sets. BLM = biotic ligand model; CA = critical accumulation; LogK = binding constant.

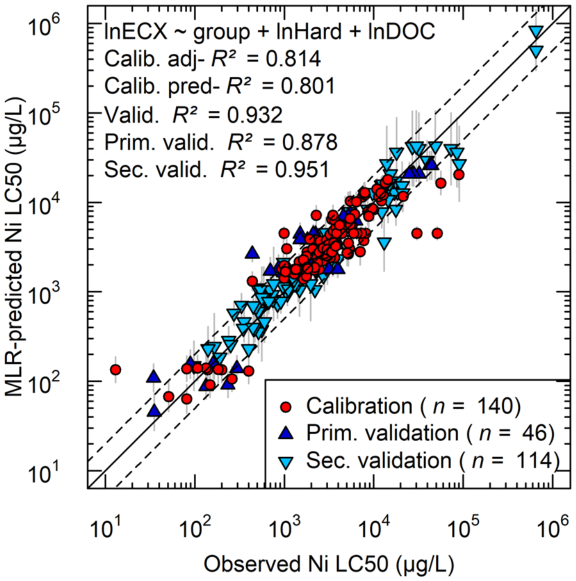

FIGURE 3: Predicted versus observed effect concentration values for the multiple linear regression model that was selected by the "step" function in R for the Pooled Acute data subset. Gray vertical lines are the 95% confidence intervals on predictions. Points between dashed lines are within a factor of 2 of perfect agreement. MLR = multiple linear regression; LC50 = 50% lethal concentration; ECX = *x*% effect concentration; DOC = dissolved organic carbon; Calib. = calibration; Valid. = validation; Prim. = primary; Sec. = secondary.

the relationships between toxicity and TMFs must be explained with a mechanistic underpinning.

Neither the primary nor the secondary validation data sets were used to fit the bioavailability relationships for either model (Figures 2 and 3). However, the primary validation data set would permit the best comparison of the models' predictive capabilities because both the bioavailability relationships and the sensitivity parameters would be fit using a different data set from the one to which they would be applied. In this case, the primary validation data set only included data for organisms, endpoints, and test types for which there were enough data to develop an MLR model, which severely limited the number of data points in the data set (see Table 1).

The secondary validation data set would allow for many more comparisons because it included additional species–endpoint–life stage groups for which there were not enough data to calibrate a species-specific model (Table 1). Using these data would require a new sensitivity parameter to be calculated from the secondary validation data set; but, because doing this would only serve to center the predictions on the measured data, the secondary validation data set would allow the fit of the bioavailability relationships of the 2 models to be evaluated (Figure 2). In the pooled models, the secondary validation data set could also be used to test the models on organisms that are not included in the calibration data set, which would test how applicable the models would be in a more real-world or regulatory context.

The performances of the 2 models using the primary and secondary validation data sets were evaluated with scoring consistent with the methods of Garman et al. (2020). The 6 scores considered evaluated various important aspects of the model, not just model fit, on a scale of 0 to 1, for a maximum total score of 6. The fit of the model was evaluated using the R^2 value as the metric for score 1 (with a minimum score of 0, even if the R^2 was negative). Scores 2 through 5 were based on the slopes of the residuals versus either the observed values or the 3 expected TMFs (i.e., hardness, DOC, and pH). The slopes were transformed so that the score would be 1 when the slope was 0, would be 0.5 when the slope was 1 or –1, and would approach 0 as the slope got steeper. This transformation was defined as

$$\text{Score}_t = 1 - \left[\frac{2}{\pi} \times \text{atan}(|m_t|)\right] \tag{5}$$

where m_t is the slope of the fitted model $\log_{10}\left(\frac{\text{ECx}_{pred}}{\text{ECx}_{obs}}\right) = (m_t \times t) + b_t$ and where t is $\log_{10}(\text{ECx}_{obs})$,

$\log_{10}$(hardness), $\log_{10}$(DOC), or pH. Score 2 was slightly different in that it was an ANCOVA slope, with the group that was used to average the sensitivity parameter serving as a categorical variable. This was used instead of an ordinary linear model because for a single organism predictions would be expected to cluster around the same measured value and be centered on a log residual of 1, which would result in most of the slopes being near zero. An ANCOVA slope allowed b_t to vary by group and produce something like an average slope for m_t, which could then be interpreted to mean how much error the model would have when predicting EC*x* values for organisms in sensitive conditions. Score 6 was the cumulative probability of predictions with an agreement factor (AF) of 2 or less. The agreement factor was calculated as follows:

$$\text{AF} = \exp\left|\ln\left(\frac{\text{EC}x_{pred}}{\text{EC}x_{obs}}\right)\right| \tag{6}$$

The cumulative probability was calculated with a Weibull plotting position ($p = i/(n+1)$), and the cumulative probability that became the score was interpolated from these 2 sets of values to where the agreement factor equaled 2. These scores were calculated and summed for the various groups of data, including for each taxonomic group, acute and chronic tests, primary and secondary validation data sets, and natural and synthetic waters.

RESULTS AND DISCUSSION

MLR model development and model selection

The stepwise regression often resulted in the same model being selected whether AIC or BIC was used as the criterion, but exceptions did occur (see Table 4). The final models were selected in each group by comparing the fit statistics between the models selected by the stepwise regression. The pooled models all included, at a minimum, DOC and hardness as parameters. All species-specific final models included hardness as a significant parameter (i.e., $p < 0.1$) except for the acute *C. dubia* model, whereas 5 of the 10 included DOC and 6 of the 10 included pH as significant parameters. The DOC parameter was only significant in one acute model (*D. pulex*) but was also present in the model for acute *D. pulicaria* with $p = 0.13$. The pH slope was significant in 3 acute models but was also present as the only parameter in the acute *C. dubia* model, with $p = 0.15$. Chronic models all included hardness as a significant parameter; DOC was a significant parameter in all but the *O. mykiss* model, and pH was a significant parameter in all but the *D. magna* reproduction model and the *P. subcapitata* model. Despite other evidence suggesting that pH is an important TMF for algae (Deleebeeck et al. 2009), the algae model in this instance did not show a strong residual trend with pH, possibly because of a difference of modeling approaches. In the development of Ni-MLR models for Australia and New Zealand, Peters et al. (2021) developed chronic models for algae, aquatic plants, invertebrates, vertebrates, and a pooled model. All of these models identified DOC and pH as TMFs; models for algae and vertebrates also identified magnesium, and the invertebrate and pooled models also identified magnesium and calcium as TMFs. The TMFs from Peters et al. (2021) notably identify DOC and pH as the major TMFs, whereas the model developed in the present study identifies DOC and hardness. Because hardness and pH tend to covary to some degree and because the Peters et al. (2021) model was specifically developed for Australian waters with often very low hardness and with species relevant for Australian guidelines, it is likely that this discrepancy is an artifact of underlying model assumptions.

Most of the final models achieved validation performance comparable to calibration performance; however, some of the species-specific models performed poorly in their primary validation data set even in their final models (Table 4). Stepwise regression did not find an acute *C. dubia* model that had all significant parameters and positive predictive R^2 and validation R^2 values. The acute *C. dubia* model in Table 4 is the best model identified by stepwise regression but would not be recommended for use without some further validation in its intended use. The model shows strong residual trends with hardness, even though it was not selected by the stepwise regression. This outcome is likely a result of too few data in the calibration data set to fit a robust model and pH acting as a surrogate for the combination of the effect of hardness on nickel toxicity and bicarbonate toxicity to *C. dubia*, as further discussed in Santore et al. (2021, in this issue). The acute *D. pulex* and chronic *O. mykiss* final models had good calibration data fit statistics and all significant slopes but did not perform well with their validation data. These 2 models only had a few primary validation data points each, and those few points the model did not predict well, so they may see better performance if there were more data available to validate against. However, unlike the acute *C. dubia* model, the acute *D. pulex* and chronic *O. mykiss* models do not show strong residual trends against any of the TMFs, so it is more likely to be a problem with the structure of the validation data giving misleading

results.

The Pooled Acute stepwise regression produced a model with only DOC and hardness, regardless of whether interactions were allowed or if the model was selected with the AIC or BIC. The model equation for the final selected Pooled Acute model was as follows:

$$\text{Acute EC}x\left(\frac{\mu\text{g}}{\text{L}}\right) = e^{0.454\times\ln(\text{hardness})+0.084\times\ln(\text{DOC})+\text{intercept}} \tag{7}$$

Figure 3 shows a comparison of predicted versus observed for the Pooled Acute model. The intercepts for the pooled models are presented in Supplemental Data 6 for this and the other pooled models.

Figure 4 shows comparisons of predicted versus observed for the Pooled Chronic models. For the Pooled Chronic model, the stepwise regression that did not allow interactions and

TABLE 4: Multiple linear regression model statistics for species-specific and pooled models: For each group of models, the preferred model is highlighted

							Slopes						
Species/duration/ endpoint	Model	Calibrated adjusted R^2	Predicted R^2	Overall validated R^2	Primary validated R^2	Secondary validated R^2	lnHard	lnDOC	pH	lnHard × lnDOC	lnHard × pH	lnDOC × pH	Intercept
Ceriodaphnia dubia 48-h LC50	All models	0.116	–0.147	–0.829	–0.829	n/a			–0.542($p =$ 0.15)				9.005
Daphnia magna 48-h LC50	No interactions (AIC)	0.246	0.186	0.434	0.434	n/a	0.196 ($p = 0.06$)		0.292 ($p = 0.003$)				5.180
	With interactions (AIC)	0.296	0.230	0.026	0.026	n/a	–2.325 ($p = 0.05$)		–1.466 ($p = 0.08$)		0.363 ($p = 0.03$)		17.319
	BIC (both)	0.208	0.145	0.124	0.124	n/a			0.359 ($p < 0.001$)				5.610
Daphnia pulex 48-h LC50	All models	0.667	0.619	–1.421	–1.421	n/a	0.750 ($p < 0.001$)	0.169 ($p < 0.001$)					5.063
Daphnia pulicaria 48-h LC50	No interactions (AIC and BIC)	0.807	0.148	0.258	0.258	n/a	0.948 ($p = 0.002$)	–0.254 ($p = 0.13$)	–0.597 ($p = 0.04$)				8.973
	With interactions (AIC and BIC)	0.839	–1.040	0.839	0.839	n/a	0.698 ($p = 0.03$)	3.453 ($p = 0.20$)	1.265 ($p = 0.34$)			–0.437 ($p = 0.17$)	–5.256
Pimephales promelas 96-h LC50 (adult)	All models	0.746	0.682	0.849	0.712	0.834	1.177 ($p < 0.001$)		–0.555 ($p < 0.001$)				8.761
Pooled Acute	All models	0.814	0.801	0.932	0.878	0.951	0.454 ($p < 0.001$)	0.084 ($p = 0.02$)					n/a
C. dubia 7-d survival and reproduction IC25	All models	0.653	0.324	0.868	0.312	0.869	0.503 ($p = 0.03$)	0.546 ($p = 0.002$)	–1.407 ($p = 0.004$)				11.390
D. magna 21-d reproduction EC50	All models	0.796	0.750	0.885	–1.073	0.902	0.510 ($p < 0.001$)	0.413 ($p < 0.001$)					2.178
D. magna 21-d survival LC50	All models	0.966	0.962	0.916	0.615	0.917	0.557 ($p < 0.001$)	0.339 ($p < 0.001$)	–0.294 ($p < 0.001$)				4.255
Oncorhynchus mykiss 17- to 26-d LC50	AIC (both)	0.821	0.709	–0.477	–0.477	n/a	0.499 ($p < 0.001$)	0.084 ($p = 0.19$)	–0.559 ($p < 0.001$)				9.216
	BIC (both)	0.810	0.702	–0.512	–0.512	n/a	0.526 ($p < 0.001$)		–0.562 ($p < 0.001$)				9.057
Pseudokirchneriella subcapitata 72-h growth EC50	All models	0.428	0.371	0.877	0.877	n/a	0.474 ($p < 0.001$)	0.221 ($p = 0.06$)					3.858

(*Continued*)

TABLE 4: (*Continued*)

Species/duration/ endpoint	Model	Calibrated adjusted R^2	Predicted R^2	Overall validated R^2	Primary validated R^2	Secondary validated R^2	Slopes						
							lnHard	lnDOC	pH	lnHard × lnDOC	lnHard × pH	lnDOC × pH	Intercept
Pooled Chronic	No interactions (AIC) and BIC (both)	0.923	0.914	0.888	0.681	0.902	0.498 ($p < 0.001$)	0.278 ($p < 0.001$)	–0.320 ($p < 0.001$)				n/a
	With interactions (AIC)	0.925	0.914	0.885	0.654	0.902	1.565 ($p = 0.02$)	–0.3275 ($p = 0.50$)	0.317 ($p = 0.38$)		–0.148 ($p = 0.10$)	0.079 ($p = 0.16$)	n/a
Pooled All	All models	0.936	0.932	0.950	0.880	0.957	0.471 ($p < 0.001$)	0.147 ($p < 0.001$)					n/a

DOC = dissolved organic carbon; LC50 = concentration lethal to 50% of a population; IC25 = concentration required for 25% inhibition of an effect; EC50 = concentration effective on 50% of a population; AIC = Akaike information criterion; BIC = Bayesian information criterion; n/a = not available.

used AIC chose a model with DOC, hardness, and pH as covariates, as did both stepwise regressions using BIC (see Figure 4). The stepwise regression that allowed interactions and used AIC was the only scenario that chose a different model, and it included the ln(hardness) × pH and ln(DOC) × pH interaction terms. Although the calculated confidence intervals were smaller for the secondary validation data with the more complex model, the validation data sets showed higher R^2 values with the simpler model with just DOC, hardness, and pH, so that model was chosen for the chronic group. The final Pooled Chronic model equation was as follows:

$$\text{Chronic EC}x\left(\frac{\mu\text{g}}{\text{L}}\right) = e^{0.498\times\ln(\text{hardness})+0.278\times\ln(\text{DOC})-0.320\times\text{pH}+\text{intercept}} \quad (8)$$

Both chronic models showed poor predictions on a number of data points. A number of *D. magna* reproduction data were among the worst-performing points, generally being underpredicted. In addition, a number of plant species were poorly performing, with a *Lemna aequinoctialis* point being the most overpredicted observation. These points all had large uncertainty bounds, which could reflect the large uncertainty that is sometimes associated with toxicity exposures and calculating EC*x* values.

The stepwise regression for the Pooled All group produced the same model whether interactions were considered or not and whether AIC and BIC were used. The model selected included only hardness and DOC (see Figure 5). The confidence intervals were fairly narrow, although some of the less accurate predictions showed larger confidence intervals. This model equation was as follows:

$$\text{EC}x\left(\frac{\mu\text{g}}{\text{L}}\right) = e^{0.471\times\ln(\text{hardness})+0.147\times\ln(\text{DOC})+\text{intercept}} \quad (9)$$

Because the Pooled All group is similar in structure to how the Santore et al. (2021, in this issue) BLM is developed and applied (not differentiating between acute and chronic or different organisms), the Pooled All model was used with the BLM in the model comparison and scoring. In addition, statistical evaluations found that there was no need to differentiate between bioavailability relationships for different organisms or acute and chronic tests. The final versions of other models developed (i.e., the species-specific models) are highlighted in Table 4. Predicted versus observed EC*x* figures for the species-specific models are included in Supplemental Data 7. Figures showing the residuals versus the observed EC*x* values and the TMFs are included in Supplemental Data 8 for all models. The pooled models were good fits for most organisms, but *C. dubia* tended to respond more to pH than did other organisms, where higher pH values were associated with lower EC*x* values, and as such the pooled model was not a particularly good fit for either the MLR or the BLM at higher pH values (pH > 8). Because *C. dubia* was at the sensitive end of the species sensitivity distribution and appeared to behave differently from most other organisms, the *C. dubia* data were further analyzed to fit a species-specific BLM to the data in Santore et al. (2021, in this issue). The pH trend, which

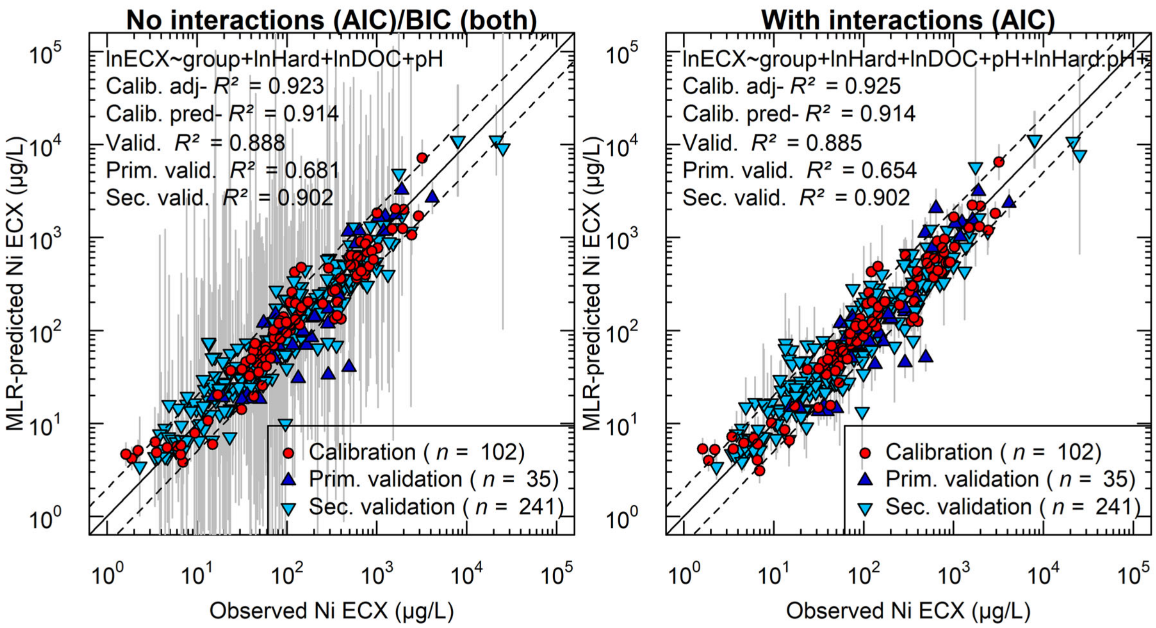

FIGURE 4: Predicted versus observed effect concentration values for the 2 MLR models that were selected by the "step" function in R for the Pooled Chronic data subset. Gray vertical lines are the 95% confidence intervals on pedictions. Points between dashed lines are within a factor of 2 of perfect agreement. AIC = Akaike information criterion; BIC = Bayesian information criterion; MLR = multiple linear regression; ECX = *x*% effect concentration; DOC = dissolved organic carbon; Calib. = calibration; Valid. = validation; Prim. = primary; Sec. = secondary.

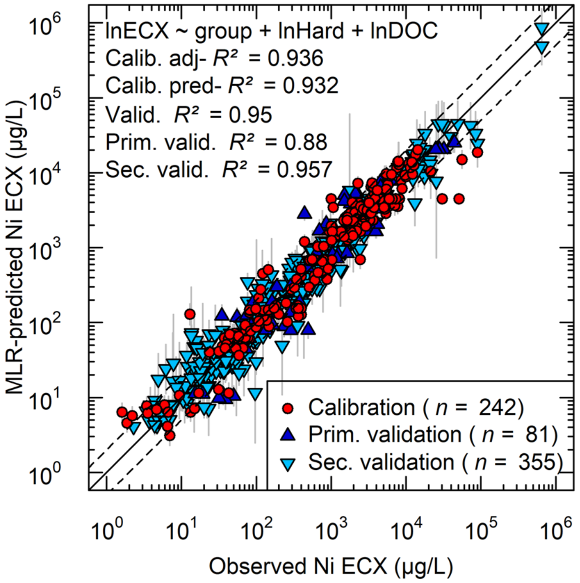

FIGURE 5: Predicted versus observed effect concentration values for the multiple linear regression model that was selected in the stepwise regression for the Pooled All data subset. Gray vertical lines are the 95% confidence intervals on predictions. Points between dashed lines are within a factor of 2 of perfect agreement. MLR = multiple linear regression; ECX = *x*% effect concentration; DOC = dissolved organic carbon; Calib. = calibration; Valid. = validation; Prim. = primary; Sec. = secondary.

appeared to be mechanistically linked to bicarbonate toxicity, is discussed in full in Santore et al. (2021, in this issue). The *C. dubia* BLM is only an approximation of a suspected mixture effect with nickel and bicarbonate; as such, it will improve the fit of the BLM to *C. dubia* but not necessarily other organisms with different bicarbonate sensitivities. As previously noted, other organisms in this database fit the pooled models quite well, but other organisms not represented in this toxicity database may be similarly affected by bicarbonate toxicity and require a similar model as *C. dubia*. Nonetheless, the pooled model fits *C. dubia* and other species reasonably well for both the MLR model and BLM because the majority of data are below pH 8, so the pooled model will be the focus of the model comparisons to follow.

BLM calculations

The same BL-binding constants were used for predictions of acute or chronic data. The critical accumulation (BLM sensitivity parameter) was calculated as the geometric mean or median of the group's individual accumulation values, which were obtained from a "speciation mode" run of the toxicity data through the BLM. The geometric mean was preferred and generally used as the group's critical accumulation value, except when it fell outside of the 25 to 75% range of the individual values, in which case the median was used. In these cases, it was assumed that outliers were skewing the geometric mean and that the median would be less influenced by outliers.

The critical accumulations, calculated for each combination of species, life stage, endpoint, and quantifier, along with the test chemistry were run through the BLM in "toxicity mode" to get predictions of ECx values, shown versus observed values in Figure 6. The confidence intervals tended to be pretty narrow, although some of the chronic data showed wider confidence intervals, spanning more than a factor of 10. The predictions that showed the largest confidence intervals tended to be

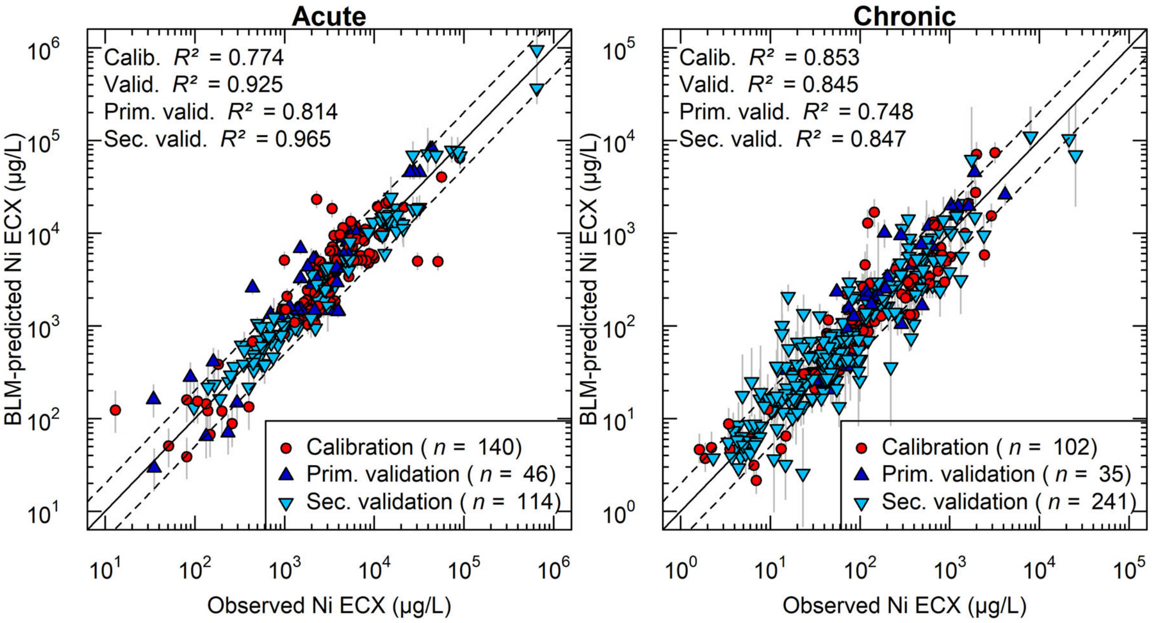

FIGURE 6: Biotic ligand model–predicted versus observed effect concentration values for the Pooled Acute data subset (left) and the Pooled Chronic data subset (right). Gray vertical lines represent the predictions from the 95% confidence limits on the critical accumulation estimates. Points between dashed lines are within a factor of 2 of perfect agreement. BLM = biotic ligand model; ECX = *x*% effect concentration; Calib. = calibration; Valid. = validation; Prim. = primary; Sec. = secondary.

those points that were least accurate. Note that both panels in Figure 6 use the same set of BLM parameters, and they are only split into acute and chronic groups to allow for a clearer picture of the information. Similar figures for other subsets of the data and residual plots are available in Supplemental Data 9 and 10.

Several outliers appeared in the BLM predictions, primarily from *C. dubia*, *Hyalella azteca*, and *P. subcapitata*. In general, outliers may stem from uncertainties related to the original toxicity data, particularly compounding the uncertainties in analytical approaches, individual organism health, or methods of calculating an effects threshold. Residual analyses of these outliers did not reveal a strong trend to any potential TMFs, so it is likely that the outliers stem from the expected uncertainties associated with all toxicological data. Overall, 86.7% of predictions on acute data and 76.2% of predictions on chronic data fall within a factor of 2 of the measured values.

Model comparisons

The BLM predictions were compared to the MLR model predictions through their predictive performance with the validation data only (i.e., excluding calibration data). Table 5 shows the derivation of the scores, consistent with Garman et al. (2020), for comparing the performance of the final Pooled All MLR model with that of the pooled BLM. The models were scored based on their R^2 (overall predictive ability); slope of residuals to observed ECx values (systematic bias on sensitivity); slope of residuals to pH, DOC, and hardness (systematic bias on bioavailability relationships); and agreement factor (consistency of predictions). These scores and others for data separated by primary and secondary validation for each species and taxonomic category are derived graphically in Supplemental Data 11 and presented in full in Supplemental Data 12. In general, the MLR model and BLM perform similarly, and either model is capable of explaining much of the variability in the observed ECx values. The following discussion focuses on the models' differences and where one performed better than the other.

The total scores for the full data set were the same: 5.37 for both the MLR model and the BLM (out of 6). The BLM performed markedly better in the residuals versus observed ECx values slope (0.74 vs 0.63 for MLR), and the MLR model performed slightly better in the slopes of residuals versus hardness and DOC (0.98 and 1.00, respectively, vs 0.96 and 0.97, respectively, for the BLM). The BLM performed slightly better in both the primary and secondary validation subsets, with total scores of 5.11 and 5.45, respectively (vs 4.96 and 5.39, respectively, for the MLR model). The slopes for residuals versus observed ECx values and residuals versus pH were the main contributors to the BLM's greater score in the case of the primary validation subset, although only the residual versus ECx was better in the case of secondary validation data. The BLM also performed slightly better in the acute data subset, with a total score of 5.56 (vs 5.38 for the MLR model), but the MLR model performed better in the chronic data subset, with a total score of 5.20 (vs 5.15 for the BLM). When subsetting the data by taxonomic category, the BLM tended to perform better with fish, whereas the MLR model tended to perform better with invertebrates, algae, and plants. The MLR model tended to perform better at the level of individual species, scoring better in 12 of the 22 species, which is a reflection of its better performance with invertebrates, the group comprising the greatest

TABLE 5: (Abbreviated) Scores comparing multiple linear regression and biotic ligand model performances using all validation data (additional comparisons provided in Supplemental Data 12): Individual scores are out of 1 and total scores are out of 6

	MLR							BLM							
Data subset	Pred vs obs R^2	Resid vs obs slope	Resid vs hard slope	Resid vs DOC slope	Resid vs pH slope	AF2 cdf	Total score	Pred vs obs R^2	Resid vs obs slope	Resid vs hard slope	Resid vs DOC slope	Resid vs pH slope	AF2 cdf	Total score	Better model
Species															
Brachionus calyciflorus	0.84	0.76	0.85	0.96	0.98	0.86	5.25	0.87	0.97	0.95	1.00	0.93	0.86	5.58	BLM
Ceriodaphnia dubia	0.72	0.56	0.95	0.95	0.98	0.74	4.90	0.68	0.64	0.79	0.96	0.87	0.75	4.69	MLR
Chironomus dilutus	0.00	0.81	0.80	0.86	0.98	0.81	4.25	0.00	0.45	0.63	0.76	0.91	0.69	3.44	MLR
Chlorella sp.	0.56	0.66	0.87	0.64	0.97	0.75	4.45	0.43	0.65	0.88	0.68	0.98	0.73	4.35	MLR
Danio rerio	0.66	0.69	0.78	0.47	0.39	0.63	3.62	0.91	0.84	0.88	0.67	0.60	0.80	4.70	BLM
Daphnia lumholtzi	0.89	0.53	0.29	0.74	0.55	0.80	3.80	0.95	0.67	0.42	0.83	0.69	0.80	4.36	BLM
Daphnia magna	0.86	0.66	0.87	0.90	0.90	0.70	4.89	0.88	0.82	0.88	0.88	1.00	0.71	5.17	BLM
Daphnia pulex	0.00	0.42	0.42	0.58	0.93	0.46	2.81	0.00	0.41	0.37	0.54	0.94	0.42	2.68	MLR
Daphnia pulicaria	0.27	0.70	0.76	0.60	0.75	0.76	3.84	0.51	0.83	0.86	0.77	0.84	0.83	4.64	BLM
Hyalella azteca	0.95	0.60	0.95	0.95	0.82	0.80	5.07	0.92	0.70	0.72	0.95	0.80	0.74	4.83	MLR
Hydra viridissima	0.82	0.75	0.79	0.87	0.88	0.83	4.94	0.67	0.72	0.72	0.97	0.91	0.67	4.66	MLR
Lampsilis siliquoidea	0.99	0.53	0.17	1.00	0.96	0.80	4.45	1.00	0.57	0.19	1.00	0.97	0.80	4.53	BLM
Lemna aequinoctialis	0	0.47	0.52	0.65	0.56	0.17	2.37	0.00	0.51	0.66	0.67	0.62	0.28	2.74	BLM
Lemna gibba	0.13	0.51	0.81	0.87	0.52	0.33	3.17	0.12	0.55	0.83	0.96	0.70	0.67	3.83	BLM
Lemna minor	0.73	0.58	0.93	0.98	0.94	0.89	5.05	0.47	0.63	0.65	0.96	0.92	0.77	4.40	MLR
Lepomis macrochirus	0.80	0.76	0.86	0.32	0.99	0.80	4.53	0.76	0.92	0.87	0.26	0.73	0.80	4.34	MLR
Lymnaea stagnalis	0.30	0.63	0.83	0.92	0.84	0.63	4.15	0.00	0.72	0.63	0.99	0.78	0.60	3.72	MLR
Melanotaenia splendida splendida	0.80	0.95	0.95	0.92	0.99	0.83	5.44	0.27	0.80	0.88	0.80	0.83	0.76	4.34	MLR
Oncorhynchus mykiss	0.88	0.69	0.99	0.88	0.96	0.83	5.23	0.83	0.80	0.90	0.81	1.00	0.74	5.08	MLR
Paracheirodon axelrodi	0.00	0.50	0.59	0.88	0.91	0.75	3.63	0.52	0.97	0.91	0.97	0.96	0.85	5.18	BLM
Pimephales promelas	0.91	0.71	0.80	0.98	0.92	0.92	5.24	0.92	0.91	0.94	0.96	0.98	0.94	5.65	BLM
Pseudokirchneriella subcapitata	0.72	0.72	0.52	0.94	0.94	0.89	4.73	0.00	0.66	0.97	0.90	0.85	0.63	4.01	MLR
Taxonomic group															
Algae	0.84	0.64	0.95	0.95	1.00	0.81	5.19	0.68	0.61	0.92	0.95	0.95	0.68	4.79	MLR
Aquatic plants	0.65	0.53	0.89	0.92	0.96	0.74	4.69	0.54	0.58	0.88	0.98	0.95	0.72	4.65	MLR
Fish	0.93	0.69	0.97	0.98	0.99	0.88	5.44	0.95	0.91	0.98	0.98	0.99	0.91	5.72	BLM
Invertebrate	0.95	0.64	0.97	0.98	0.99	0.81	5.34	0.94	0.72	0.95	0.94	0.96	0.80	5.31	MLR
Test type															
Acute	0.92	0.67	0.99	0.94	0.99	0.87	5.38	0.93	0.84	0.97	0.95	0.99	0.88	5.56	BLM
Chronic	0.88	0.62	0.99	0.95	0.98	0.78	5.20	0.84	0.67	0.95	0.99	0.95	0.75	5.15	MLR
Validation type															
Primary	0.88	0.65	0.87	0.95	0.90	0.71	4.96	0.87	0.77	0.86	0.94	0.99	0.68	5.11	BLM
Secondary	0.96	0.63	0.99	0.99	0.98	0.84	5.39	0.95	0.71	0.98	1.00	0.98	0.83	5.45	BLM
All data	0.95	0.63	0.98	1.00	1.00	0.81	5.37	0.94	0.72	0.96	0.97	0.98	0.80	5.37	

MLR = multiple linear regression; BLM = biotic ligand model; Pred = predicted; obs = observed; Resid = residual; Hard = hardness; DOC = dissolved organic carbon; AF2 cdf = agreement factor of 2 cumulative distribution function.

number of species. The MLR performed much better with *Melanotaenia splendida splendida*, whereas the BLM was much better in *Paracheirodon axelrodi* and *Danio rerio*, and 9 of the 22 species performed about the same for both models.

The BLM performed much better than the MLR model with score 2, the slope of residuals versus observed EC*x* values. The slopes used to calculate the score were negative most of the time (for both models), but the slope in the BLM score was generally shallower than that in the MLR model's score, meaning that the BLM would have less tendency to overestimate more sensitive EC*x* values and underestimate less sensitive EC*x* values. The species for which the MLR score was significantly better than the BLM score included a *Chironomus dilutus*, *M. splendida splendida*, and *P. subcapitata*.

The BLM also performed somewhat better with score 5, the slope of residuals versus pH. Because the Pooled All MLR model chosen for comparison did not include pH effects, this was somewhat to be expected. Surprisingly, the MLR model performed better for *C. dubia*, despite the fact that the Pooled All MLR model ignored any effect of pH, and the species-specific *C. dubia* MLR models featured a large pH slope—exclusively so for the acute *C. dubia* model. The MLR model often scored better than the BLM for score 1 (predicted vs observed R^2) and scores 3 and 4 (residual vs hardness or vs DOC slopes, respectively). Because the Pooled All MLR had DOC and hardness as its 2 explanatory variables and the fitting of an MLR model is based on minimizing a sum of squared errors (i.e., maximizing the R^2), it makes sense that the MLR model would perform somewhat better than a mechanistic model for these scores. The MLR tended to perform slightly better on score 6 (agreement factor of <2 cumulative probability).

Although the confidence intervals are not included in the scoring, they can be a good way to informally compare the uncertainty associated with each model. Both models show higher uncertainty around the predictions of points that are less accurate (further from the 1:1 line).

The MLR equations can help inform water quality guidelines by providing a simple means for including TMFs. The comparison in the present study shows that the MLR model, when it is informed by a large amount of good-quality data, can provide information about nickel bioavailability and that its performance is on par with the mechanistic BLM. The availability of simple approaches for improving guidelines may also help harmonize guidelines in different jurisdictions around the world. Currently, the US guideline is fairly outdated and is based on a simple hardness equation (US Environmental Protection Agency 1996) that does not consider DOC, which this analysis recognizes as an important TMF. Canada has recently started looking at developing a bioavailability-based approach for its nickel guideline, and that approach will likely consider both hardness and DOC. In the European Union, a bioavailability model is used that incorporates DOC, calcium, and pH. Peters et al. (2018) used Australian ecotoxicity tests to show that bioavailability principles are supported for water chemistry ranges observed in Australia in natural water experiments; their findings provided the basis for establishing bioavailability-based water quality guidelines in Australia and New Zealand (Peters et al. 2021; Stauber et al. 2021). To some extent, differences in guidelines are the result of the organisms (e.g., native species) selected to inform the guidelines.

CONCLUSIONS

As shown in the present study, MLR models can be developed for nickel, and these models account for the same TMFs considered by mechanistic models like the BLM. The stepwise regressions provided fairly consistent results, whether or not they allowed interaction terms in the MLR models and whether they used the AIC or BIC statistic. The interaction terms were not always selected, even when allowed; and they never produced a model that was deemed best overall when applied to independent validation data. A pooled model approach resulted in fairly consistent predictions and allowed for flexibility when examining relatively unstudied species so that a species-specific model could be developed.

When applying the overall pooled MLR model to independent validation data and comparing its performance with that of a pooled mechanistic model (i.e., the BLM), both models performed well. The BLM performed slightly better than the MLR model in the primary validation data, when there were enough data available to calibrate a sensitivity parameter. The models performed equally well in the secondary data, where the sensitivity parameter was set to the data; this approach centered any predictions but still tested the ability of the model to capture the effect of TMFs. The MLR model performed better than the BLM for invertebrates, which had the most calibration data available, and was better at more accurately capturing the slopes for the TMFs for which it had parameters. These various strong points for each model would lend themselves to different applications, so both models are useful in research and regulatory applications, depending on the particular needs of the task.

Supplemental Data—The Supplemental Data are available on the Wiley Online Library at https://doi.org/10.1002/etc.5063.

Acknowledgment—The present study was performed with the generous support of the Nickel Producers Environmental Research Association.

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

REFERENCES

AECOM Technical Services. 2011. Acute toxicity of selected mixtures of nickel, cobalt, copper, and iron to freshwater algae (*Pseudokirchneriella subcapitata*). Fort Collins, CO, USA.

Allen DM. 1974. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* 16:125–127.

Alsop D, Wood CM. 2011. Metal uptake and acute toxicity in zebrafish: Common mechanisms across multiple metals. *Aquat Toxicol* 105:385–383.

ASTM International. 1996. Standard guide for conducting acute toxicity tests on test materials with fishes, macroinvertebrates, and amphibians. E 729-96. In *Annual Book of ASTM Standards*, Vol 11.06. Philadelphia, PA.

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

Canadian Council of Ministers of the Environment. 2007. Canadian water quality guidelines for the protection of aquatic life: A protocol for the derivation of water quality guidelines for the protection of aquatic life 2007. Winnipeg, MB, Canada.

DeForest D, Santore R, Ryan A, Church B, Chowdhury M, Brix K. 2017. Development of biotic ligand model–based freshwater aquatic life criteria for lead following US Environmental Protection Agency guidelines. *Environ Toxicol Chem* 36:2965–2973.

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

Deleebeeck NME, De Schamphelaere KAC, Heijerick DG, Bossuyt BTA, Janssen CR. 2008a. The acute toxicity of nickel to *Daphnia magna*: Predictive capacity of bioavailability models in artificial and natural waters. *Ecotoxicol Environ Saf* 70:67–78.

Deleebeeck NME, De Schamphelaere KAC, Janssen CR. 2007. A bioavailability model predicting the toxicity of nickel to rainbow trout (*Oncorhynchus mykiss*) and fathead minnow (*Pimephales promelas*) in synthetic and natural waters. *Ecotoxicol Environ Saf* 67:1–13.

Deleebeeck NME, De Schamphelaere KAC, Janssen CR. 2008b. A novel method for predicting chronic nickel bioavailability and toxicity to *Daphnia magna* in artificial and natural waters. *Environ Toxicol Chem* 27:2097–2701.

Deleebeeck NME, De Schamphelaere K, Janssen CR. 2009. Effects of Mg^{2+} and H^+ on the toxicity of NI^{2+} to the unicellular green alga *Pseudokirchneriella subcapitata*: Model development and validation with surface waters. *Sci Total Environ* 407:1901–1914.

De Schamphelaere K, Van Laer L, Deleebeeck N, Muyssen B, Degryse F, Smolders E, Janssen C. 2006. Nickel speciation and ecotoxicity in European natural surface waters: Development, refinement and validation of bioavailability models. Final report. Laboratory for Environmental Toxicology and Aquatic Ecology, Ghent University, Gent, Belgium.

Environment and Climate Change Canada. 2019. Federal environmental quality guidelines. Copper. Gatineau, QC, Canada.

Esbaugh AJ, Mager EM, Brix KV, Santore R, Grosell M. 2013. Implications of pH manipulation methods for metal toxicity: Not all acidic environments are created equal. *Aquat Toxicol* 130–131C:27–30.

European Commission. 2011. Common implementation strategy for the Water Framework Directive (2000/60/EC). Guidance document 27. Technical guidance for deriving environmental quality standards. Brussels, Belgium.

Ferreira C, Olivera F, Djokic D. 2010. Development of large topographic data sets for hydrologic analysis. *Proceedings*, GeoShanghai 2010 International, Shanghai, China, June 3–5, 2010, pp 2772–2780.

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.

Hirsch RM, De, Cicco LA. 2015. User guide to exploration and graphics for RivEr Trends (EGRET) and dataRetrieval: R packages for hydrologic data (Ver 2.0). Chapter 10, Section A, Statistical analysis. Book 4, Hydrologic analysis and interpretation. US Geological Survey, Reston, VA.

Hopper T. 2014. Can we do better than R-squared? *Competitive Organizations Through High-Performance Learning* (blog). 2014 May 16. [cited 2019 August 7]. Available from: https://tomhopper.me/2014/05/16/can-we-do-better-than-r-squared/

Keller AE, Zam SG. 1991. The acute toxicity of selected metals to the freshwater mussel, *Anodonta imbecilis*. *Environ Toxicol Chem* 10:539–546.

Kim D, Chae Y, An Y-J. 2017. Mixture toxicity of nickel and microplastics with different functional groups on *Daphnia magna*. *Environ Sci Technol* 51:12852–12858.

Kozlova T, Wood CM, McGeer JC. 2009. The effect of water chemistry on the acute toxicity of nickel to the cladoceran *Daphnia pulex* and the development of a biotic ligand model. *Aquat Toxicol* 91:221–228.

Meyer JS, Santore RC, Bobbit JP, Debrey LD, Boese CJ, Paquin PR, Allen HE, Bergman HL, Ditoro DM. 1999. Binding of nickel and copper to fish gills predicts toxicity when water hardness varies, but free-ion activity does not. *Environ Sci Technol* 33:913–916.

Nys C, Janssen CR, Van Sprang P, De, Schamphelaere KAC. 2016. The effect of pH on chronic aquatic nickel toxicity is dependent on the pH itself: Extending the chronic nickel bioavailability models. *Environ Toxicol Chem* 35:1097–1106.

Parametrix. 2005. Acute toxicity of nickel at elevated hardness, alkalinity and pH. Prepared for Nickel Producers Environmental Research Association. Albany, OR, USA.

Pavlaki MD, Pereira R, Loureiro S, Soares AMVM. 2011. Effects of binary mixtures on the life traits of *Daphnia magna*. *Ecotoxicol Environ Saf* 74:99–110.

Peters A, Merrington G, Schlekat C, De Schamphelaere K, Stauber J, Batley G, Harford A, van Dam R, Pease C, Mooney T, Warne M, Hickey C, Glazebrook P, Chapman J, Krassoi R. 2018. Validation of the nickel biotic ligand model for locally relevant species in Australian freshwaters. *Environ Toxicol Chem* 37:2566–2574.

Peters A, Merrington G, Stauber J, Golding L, Batley G, Gissi F, Adams M, Binet M, McKnight K, Schlekat CE, Garman E, Middleton E. 2021. Empirical bioavailability corrections for nickel in freshwaters for Australia and New Zealand water quality guideline development. *Environ Toxicol Chem* 40:113–126.

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

Sakamoto Y, Ishiguro M, Kitagawa G. 1986. *Akaike Information Criterion Statistics*. D. Reidel, Dordrecht, The Netherlands.

Santore RC, Croteau K, Ryan AC, Schlekat C, Middleton E, Garman E. 2021. A review of water quality factors that affect nickel bioavailability to aquatic organisms: Refinement of the biotic ligand model for nickel in acute and chronic exposures. *Environ Toxicol Chem* 40:2121-2134.

Santore RC, Mathew R, Paquin PR, DiToro D. 2002. Application of the biotic ligand model to predicting zinc toxicity to rainbow trout, fathead minnow, and *Daphnia magna*. *Comp Biochem Physiol C Toxicol Pharmacol* 133:271–285.

Schubauer-Berigan MK, Dierkes JR, Monson PD, Ankley GT. 1993. pH-dependent toxicity of Cd, Cu, Ni, Pb, and Zn to Ceriodaphnia dubia, Pimephales promelas, Hyalella azteca and Lumbriculus variegatus. *Environ Toxicol Chem* 12:1261–1266.

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. *Environ Toxicol Chem* 40:100–112.

Tarpey T. 2000. A note on the prediction sum of squares statistic for restricted least squares. *Am Stat* 54:116–118.

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

US Environmental Protection Agency. 1994. Short-term methods for estimating the chronic toxicity of effluents and receiving waters in freshwater organisms, 3rd ed. EPA 600/4-91-002. Cincinnati, OH.

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, Washington, DC.

US Environmental Protection Agency. 2007. Aquatic life ambient freshwater quality criteria—Copper. EPA-822-R-07-001. Washington, DC.

van Genderen E, Stauber JL, Delos C, Eignor D, Gensemer RW, McGeer J, 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.

Water Environment Research Federation, Wu KB, Paquin PR, Navab V, Mathew R, Santore RC, Di Toro DMPD. 2003. Development of a biotic ligand model for nickel: Phase I. Alexandria, VA, USA.

Windward Environmental. 2017. Biotic Ligand Model Windows® Interface, Research Ver 3.16.2.41: User's Guide and Reference Manual. Seattle, WA, USA.