# Simplified random forest models predict reference-condition water chemistry as well as more complex models

**Daniel Nelson**[1,2], **Jennifer L. Courtwright**[1,3], **and Charles P. Hawkins**[1,4]

[1]Department of Watershed Sciences, National Aquatic Monitoring Center, and Ecology Center, Utah State University, Logan, Utah, USA

**Abstract:** Elevated nutrient concentrations and increased salinization threaten the ecological integrity of freshwater habitats worldwide. Many waterbodies are experiencing continued or worsening water quality despite decades of monitoring and remediation efforts. Understanding the extent to which water quality has been impaired requires that we compare observed water-quality measurements with naturally occurring benchmark values. Several approaches have been used to estimate these benchmarks, but many of these approaches provide only a single value for a given region, which will typically either under- or overestimate reference conditions at individual sites. Predictive models for estimating site-specific reference conditions exist for some water-quality indicators, but their performance in terms of accuracy, precision, or coverage can limit their use. Furthermore, models can be difficult to implement if they rely on predictors that are not readily available. In this study, we attempted to improve existing random forest models used to predict naturally occurring, site-specific spatial variation in levels of specific conductivity (SC) and concentrations of total N (TN) and total P (TP). We predicted that we could improve model performance and ease of implementation by training models on larger datasets and using predictor variables from a common, nationally available dataset. We compared predictions from the revised models with predictions made by the original set of models at both reference and test sites. In addition, we compared 2 methods of estimating upper prediction limits that could be used to set site-specific benchmarks and compared these site-specific benchmarks with regional benchmarks. The performance of the revised SC, [TN], and [TP] models were similar to that of the original models, but the ease of implementation was greatly improved through the use of a nationally available dataset of watershed-scale predictors. Site-specific and regional benchmarks differed considerably, with regional benchmarks being higher for SC and lower for TN and TP than site-specific benchmarks derived from the models. Our results suggest that site-specific predictive water-chemistry models based on easily obtainable predictors from a nationally available dataset can perform as well as those based on predictors that require more advanced geographic information system analysis.

**Key words:** random forests, specific conductivity, total nitrogen, total phosphorus, water chemistry, benchmarks, prediction intervals, streams, rivers

## INTRODUCTION

Water-quality degradation has profoundly altered the ecological integrity of streams and rivers across the globe (Meybeck 2004, Lintern et al. 2018, Akhtar et al. 2021). For decades, the United States and other countries have used ecological (e.g., water quality and aquatic biota) assessments to inform the management of freshwater ecosystems (Keiser and Shapiro 2019). However, many countries are still experiencing continued or worsening water quality despite decades of monitoring and remediation (Schwarzenbach et al. 2010, Stets et al. 2020). Understanding the extent to which water quality has been impaired requires that we assess whether observed conditions differ from the range of natural variation

Email addresses: [2]dann461@ecy.wa.gov; Present address: Washington State Department of Ecology, Olympia, Washington, USA; [3]jennifer.courtwright@usu.edu; [4]chuck.hawkins@usu.edu

**ORCID iDs**: D. Nelson, https://orcid.org/0000-0001-5263-6308; J. Courtwright, https://orcid.org/0000-0002-9864-8547; C. Hawkins, https://orcid.org/0000-0003-1247-0248

Received 21 June 2024; Accepted 2 May 2025; Published online 16 June 2025; Associate Editor, Matthew Baker

*Freshwater Science*, volume 44, number 3, September 2025. © 2025 The Society for Freshwater Science. All rights reserved. Published by The University of Chicago Press for the Society for Freshwater Science. https://doi.org/10.1086/736692

expected to occur in the absence of human-caused alteration (Hawkins et al. 2010, Soranno et al. 2011).

Many assessment programs use minimally or least-disturbed reference sites to establish benchmarks for inferring the condition of other sites (Stoddard et al. 2006). These benchmarks are often derived at regional scales (i.e., a common benchmark is applied everywhere within a region; USEPA 2000, Suplee et al. 2007, Herlihy and Sifneos 2008), but site-specific benchmarks are more effective in distinguishing human-caused alterations from naturally occurring variation, thus minimizing type I and type II errors of inference (Hawkins et al. 2010, Ohlendorf et al. 2011, Olson and Hawkins 2013, van Dam et al. 2017). Regional benchmarks do not account for environmental conditions (e.g., climate, lithology, chemistry) that can vary among sites within regions and affect water chemistry. Moreover, benchmarks should be set at the scale at which management is applied—individual water bodies. Site-specific benchmarks are especially needed in heterogeneous regions where application of single, regional benchmarks is likely to be either under- or overprotective. Regional benchmarks often lack the precision and accuracy necessary to set appropriate numerical criteria for water-quality constituents (Hawkins et al. 2010, van Dam et al. 2017). Thus, approaches are needed to easily set site-specific benchmarks based on readily available environmental data.

Site-specific benchmarks are typically set by developing predictive models from observations made at a series of reference-quality sites (Hawkins et al. 2010, van Dam et al. 2019). Models based on machine learning algorithms (e.g., random forests, boosted regression trees) appear to be especially well suited to such tasks (Lek et al. 1999, Shen et al. 2020, Zhu et al. 2022, Yan et al. 2024). For example, random forest models have been developed to predict site-specific background levels of specific conductivity (SC) (Olson and Hawkins 2012, Le et al. 2019, Olson and Cormier 2019) and concentrations of total N (TN) and total P (TP) (Olson and Hawkins 2013) based on variation in a set of naturally occurring landscape features (e.g., catchment slope, soil erodibility) obtained from geographic information systems (GIS). Site-specific predictions derived from these models usually represent an improvement over regional benchmark approaches, but the predictions can still be imprecise (particularly for [TN] and [TP]), limited to sites within a particular range of naturally occurring conditions (e.g., applicable only to sites within a specific elevation range), or derived from models that use temporally dynamic predictor variables. In the latter case, the use of temporally dynamic variables can lead to shifting baseline conditions (i.e., when reference conditions are adjusted for the effects of a human-caused factor such as climate change rather than anchored at a standard period in the past). In addition, some of the predictor variables that these models require are difficult to calculate by anyone other than advanced GIS users and are sometimes not reproducible because GIS software is not designed to record all of the tasks the original developer performed. Consequently, there is a need for temporally anchored models that can accurately and precisely predict reference conditions for SC, [TN], and [TP] across a wider range of naturally occurring site conditions.

Predictions of site-specific reference conditions have little utility if their accuracy and precision are not known. To establish useful site-specific benchmarks, model predictions should include measures of prediction uncertainty (Olson and Hawkins 2013). Such uncertainty can arise from a myriad of sources, such as error in measuring predictor values (e.g., site elevation, watershed area), error in measuring the response variable (e.g., [TN] and [TP]), and imperfect model structure (i.e., the model inaccurately represents the response–predictor relationship). Prediction uncertainty has often been quantified by estimating prediction intervals (PIs) (Gibbons 1987, Olson and Hawkins 2013, Zhou et al. 2022), the upper limit (PL) of which represents the highest probable naturally occurring water-chemistry concentration at a site. Several methods for estimating PIs have been proposed, but no consensus exists regarding how PIs should be calculated for random forest models. This lack of consensus adds further uncertainty to identifying the upper limits of expected conditions for such models.

Our main objective was to improve the random forest models currently being used by some federal and state agencies in the western United States (hereafter referred to as the original models) to predict reference conditions for SC (Olson and Hawkins 2012, Olson and Cormier 2019), [TN], and [TP] (Olson and Hawkins 2013). Second, we wanted to assess how strongly the method of calculating PIs affected inferences of water-quality impairment. Third, we wanted to determine whether site-specific benchmarks were less or more protective of individual waterbodies than regionally derived benchmarks such as those used by the United States Environmental Protection Agency's (EPA) National Rivers and Streams Assessment (NRSA). Fourth, we expected that we could improve model performance and broaden the applicability of the original models to a larger environmental space than covered by the original set of reference sites by retraining the models on an expanded set of reference sites. Fifth, we expected that we could improve the ease of conducting assessments (i.e., eliminate the need for advanced GIS analyses) by using reproducible predictor variables contained in an easily accessible and nationally available dataset. Sixth, we wanted to alleviate shifting-baseline issues by using static predictor variables that temporally anchored reference conditions to a fixed window of time. Last, we expected that regional benchmarks would be either over- or underprotective relative to site-specific benchmarks and thus more prone to type I and II errors than site-specific benchmarks.

## METHODS

### Reference and test-site selection

We compiled observations of SC, [TN], and [TP] from a network of reference sites distributed across the western United States. Candidate reference sites were initially identified by the agency that sampled the sites (e.g., United States Geological Survey [USGS], EPA, state agencies). These candidate sites were later screened to ensure that their respective watersheds were least disturbed (sensu Stoddard et al. 2006) by human activity (Olson and Hawkins 2012, 2013, Olson and Cormier 2019). Sites were first screened to verify that their catchments had <10% agriculture or urban land use (Olson and Hawkins 2013). In addition, aerial photographs and maps were inspected for other evidence of human impacts (e.g., mines, ranches, clearcuts). We further screened the original sets of 1391, 665, and 752 reference sites from Olson and Hawkins (2012, 2013) used to model SC, [TN], and [TP], respectively, to minimize potential problems of data independence. Some of the original reference-site observations occurred on the same stream segment, so we selected just 1 sample from each segment when retraining models. After removing duplicate sites, we were left with 1359, 617, and 736 of the original observations used in the SC, TN, and TP models, respectively. However, we were able to add additional reference sites for all 3 models by including sites from Olson and Cormier (2019) and other reference-quality sites from across the western United States (J. R. Olson, California State University, Monterey Bay, California, personal communication). As a result, we were able to train the revised SC, TN, and TP models on 1912, 699, and 966 observations (Fig. 1A–C). New reference-site observations for the SC model were located in every western state (Fig. 1A), but new TN data were available only from Montana (Fig. 1B). For the revised TP model, we were able to add new observations from Oregon, Nevada, Utah, and Montana (Fig. 1C).

In addition to selecting reference sites for model calibration, we selected >1500 test sites in the western United States to evaluate the effects of modeling decisions on inferences of water-quality condition. All test sites were sampled as part of the Bureau of Land Management's (BLM) Assessment, Inventory and Monitoring (AIM) Strategy (Toevs et al. 2011). The AIM program uses standardized and consistent methods to assess the condition of natural resources on BLM lands. The AIM test sites included sites sampled as part of spatially balanced survey designs plus some targeted locations of specific management interest to the BLM. Sites included both wadeable and nonwadeable perennial stream and river reaches represented in the National Hydrography Dataset (NHD) (BLM 2015). Not all test sites had observations for all 3 water-quality constituents, but we were able to identify 1957, 1539, and 1547 test sites with SC, TN, and TP data, respectively.

### Specific conductivity and nutrient concentration data

SC (normalized to 25°C), [TN], and [TP] at reference sites were taken directly from the Olson and Hawkins (2012, 2013) and Olson and Cormier (2019) datasets or downloaded from the National Water Quality Monitoring Council Water Quality Portal (NWQMC et al. 2021; https://www.waterqualitydata.us). Reference-site data for

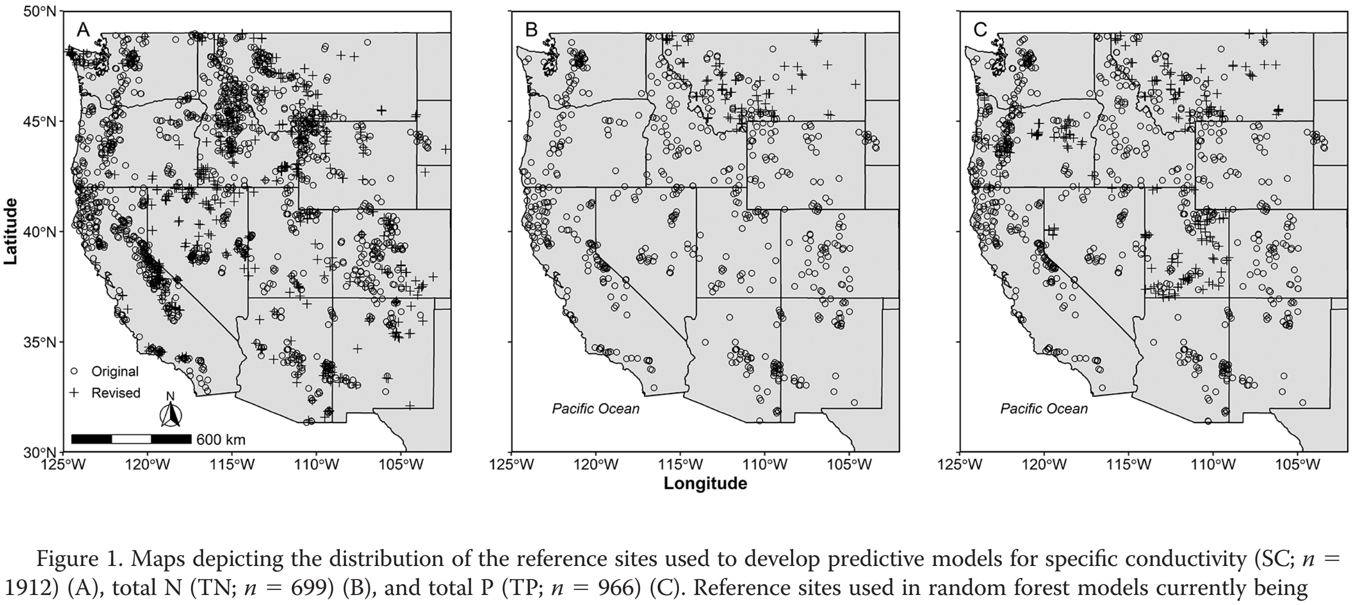

Figure 1. Maps depicting the distribution of the reference sites used to develop predictive models for specific conductivity (SC; $n = 1912$) (A), total N (TN; $n = 699$) (B), and total P (TP; $n = 966$) (C). Reference sites used in random forest models currently being used by some federal and state agencies in the western United States (original models) are shown as open circles, and newly added sites included in retrained models (revised models) are shown as plus signs. Reference-site data for SC ranged from 1965 to 2016, and data for TN and TP ranged from 1973 to 2015 and 1973 to 2019, respectively.

SC ranged from 1965 to 2016, and data for TN and TP ranged from 1973 to 2015 and 1973 to 2019, respectively. SC was measured in the field or laboratory, depending on the sampling agency. [TN] at the original set of reference sites was measured from persulfate digestion and colorimetry or the sum of Kjeldahl [N], [$NO_3^-$], and [$NO_2^-$] estimates. For the added reference sites, we only used [TN] measured from persulfate digestion and colorimetry because of uncertainty about what other methods were used and reported in the data portal. [TP] was measured from unfiltered water samples following persulfate digestion and colorimetry. We ensured that all data were standardized to the same units (μS/cm for SC and μg/L for [TN] and [TP]).

We assigned all reference-site observations to their associated stream segments in the NHD Plus V2 (NHDPlus V2; McKay et al. 2012). Each stream segment has a unique common identifier called a COMID. To assign COMIDs to sites, we spatially joined sample locations (i.e., points) to NHDPlus V2 catchment polygons for each stream segment. If >1 sample was collected from the same COMID segment, we randomly selected 1 of the sites for analysis and dropped the other sample(s). However, we checked the randomly selected samples to ensure that concentrations fell within an expected range of values for that site. For example, if a site was sampled 5× over several years and 4 of those 5 sample concentrations were similar but 1 was much different than the other 4, we randomly selected a sample from 1 of the 4 similar samples because these values were most likely to represent typical background concentrations. If [TN] or [TP] measurements were below the limits of detection, we used the limit of detection as the final concentration. Limits of detection differed by agency or analyzing laboratory and ranged from 1 to 10 μg/L for [TP] and 5 to 60 μg/L for [TN]. There was no limit of detection for SC. Spatially modeling water-chemistry data with multiple detection limits across regions can potentially affect model performance and bias predictions (Fu and Wang 2011). Therefore, we explored whether replacing all limit-of-detection values with the highest limit-of-detection value in the combined datasets (e.g., a [TP] value of 3 μg/L replaced with a value of 10 μg/L) would affect model performance. There was no appreciable difference in performance between models that used these rolled-up limit-of-detection values and those that used the original values. However, models that were trained on the original data were more accurate in predicting lower concentrations than those that used rolled-up values. We therefore chose to use the original, variable limit-of-detection data when developing the models. Sites at which nutrient concentrations were at the limits of detection (e.g., 1–10 μg/L for TP) were distributed throughout the study area but did show some bias with regard to geography. For example, samples collected from the state of Washington had a relatively high proportion of measurements below the limit of detection (Fig. S1).

We downloaded publicly available data for test sites from the BLM National AIM Lotic Indicators Hub (https://gbp-blm-egis.hub.arcgis.com/pages/aim), covering the years 2013 to 2021. Specific conductivity was measured by BLM personnel in the field with YSI sondes (Yellow Springs Instruments, Yellow Springs, Ohio) at each site. TN and TP samples were processed by the Utah State University (USU) Aquatic Biogeochemistry Laboratory (Logan, Utah). [TN] was measured from unfiltered grab samples following a potassium persulfate digestion and a cadmium reduction. The USU lab's limit of detection for [TN] was 12 μg/L. [TP] was measured from unfiltered grab samples following potassium persulfate digestion, an ascorbic acid molybdenum reaction, and colorimetric analysis. The USU lab's limit of detection for [TP] was 15 μg/L.

## Candidate predictor variables

We appended StreamCat (https://www.epa.gov/national-aquatic-resource-surveys/streamcat-dataset#access-streamcat-data; Hill et al. 2016) variables to sites based on NHD COMID identifiers. First, we used the *StreamCatTools* package (Weber et al. 2024) in R (version 4.2.0; R Project for Statistical Computing, Vienna, Austria) to append all available StreamCat variables to sites. Before modeling, we excluded all catchment-level (sensu Hill et al. 2016) variables because initial analyses indicated that watershed-level predictors performed better than catchment-level predictors. Hill et al. (2016) define a catchment as the portion of the watershed laterally adjacent to the stream reach that drains directly into the reach. Next, we eliminated StreamCat variables that characterize anthropogenic impacts (e.g., dam density, coal mine density, road density) because we were only interested in modeling natural variation among sites. Finally, we eliminated variables that likely had little influence on SC and nutrient concentrations in streams (e.g., predicted wetted width of a stream). In addition to StreamCat variables, we included the day of the year (DOY) on which samples were collected for TN and TP models because nutrient concentrations can vary seasonally. We were unable to calculate DOY for the SC model because the sampling date was missing for most of the original reference sites used by Olson and Hawkins (2012). Overall, we ended up with 45 candidate predictor variables for the SC model and 46 candidate predictors for the TN and TP models (see Table S1 for a complete list and description of candidate predictor variables).

## Model development and performance

We used random forest modeling (Breiman 2001) to predict spatial variation in reference-condition levels of SC, [TN], and [TP] at the revised reference sites. First, we used the vsurf procedure (with default settings) in the *VSURF* package (version 1.2.0; Genuer et al. 2015) in R to select a parsimonious and interpretable set of predictors for each model. The vsurf procedure is a stepwise feature-selection method based on random forests with the purpose of removing redundant predictors from a dataset. It assesses variable importance from the increase in mean square error (MSE) when a given variable is permuted. The importance of the variable increases as the difference in MSE between the model using the unpermuted variable and the model using the permuted variable increases. Overall, an increase in MSE indicates a decrease in model performance. Next, we implemented random forest modeling with the R package *randomForest* (version 4.7-1.1; Liaw and Wiener 2002) set to default settings. For each water-chemistry constituent, we included those predictors identified by the vsurf procedure as being important. We then used the random forest out-of-bag observations to assess model performance in terms of $r^2$ values, the Nash–Sutcliffe model efficiency coefficient (NSE), root mean square error (RMSE), and mean absolute error (MAE) from the linear regression of observed values on predicted values. NSE is a measure of the correspondence between observations and their predictions. If NSE = 1, the predictions perfectly correspond with the observations, whereas if NSE = 0, the model has the same explanatory power as the mean of all observations. When NSE < 0, model performance is worse at making predictions than the mean of observations. RMSE is a measure of the average distance between a model's predicted values and the observations and is calculated as the SD of model residuals. MAE is a measure of how close the predictions are to the observations. Both RMSE and MAE are useful in evaluating a model because they are computed and reported in the same units as the dependent variable. RMSE and MAE are also negatively oriented metrics, meaning that lower values indicate better model performance than higher values.

We compared the performance of the original and revised models for both reference and test sites. First, we obtained model predictions for the original set of reference sites used by Olson and Hawkins (2012, 2013) and Olson and Cormier (2019). We were not able to make predictions for the added reference sites with the original models because of our inability to reliably duplicate some predictor values included in those models (e.g., mean channel slope for the original TP model). Therefore, our comparisons of revised vs original model predictions for reference sites are limited to the original sets of reference sites. Next, we appended the StreamCat variables selected for each model to test sites based on NHD COMID identifiers and used the revised models to make predictions of SC, [TN], and [TP] at test sites. The BLM's AIM program has used the original predictive models (Olson and Hawkins 2012, 2013) to set site-specific benchmarks (i.e., reference conditions) for SC, [TN], and [TP] across the western United States. Therefore, predicted reference values for SC, [TN], and [TP] based on the original models were already available for many of the test sites. We used Pearson's correlation coefficient to assess the association between original model predictions and revised model predictions. We then used reduced major axis regression to regress the original model predictions ($y$-axis) against the revised model predictions ($x$-axis). Based on the reduced major axis regression results, we assessed whether the slope of the regression line was equal to 1 based on the 95% CIs around the slope. A slope <1 indicates that, on average, the revised model overpredicts relative to the original model at lower values and underpredicts at higher values. A slope >1 indicates that the revised model generally underpredicts relative to the original model at lower values and overpredicts at higher values.

### Identifying sites outside reference-site environmental space

Extrapolating beyond the environmental conditions represented in the training dataset can lead to misleading inferences. To mitigate this risk, we assessed whether each test site fell within the naturally occurring environmental space defined by the reference sites. To identify test sites with environmental characteristics that were outside the environmental space defined by the reference sites, we used the nearest-neighbor approach developed by Vander Laan and Hawkins (2014). Briefly, we chose 5 environmental variables that broadly characterized spatial variation in naturally occurring environmental conditions across the study region: watershed area (WsAreaSqKm), elevation (ElevWs), maximum temperature (Tmax8110Ws; 30-y normal maximum temperature), minimum temperature (Tmin8110Ws; 30-y normal minimum temperature), and precipitation (Precip8110Ws; 30-y normal mean precipitation). See Table S1 for descriptions of each variable. We then standardized the 5 variables by scaling between minimum and maximum values observed at reference sites. We calculated multivariate Euclidean distances between each reference site and all other reference sites based on the standardized values. Next, we calculated the mean Euclidean distance of each reference site to the 10 nearest reference sites and used the 90th percentile of this distribution as a threshold for defining whether a test site was outside reference-site environmental space. We applied this test to all candidate test sites by calculating the mean distance of each test site to the 10 nearest reference sites, and we flagged a test site as an outlier if the mean distance exceeded the 90th percentile threshold. We removed test sites flagged as outliers from the test dataset before comparing model predictions from the original and revised models.

To determine if the revised models can be applied to a larger number of test sites than the original models, we performed the above procedure based on both the revised set of reference sites and the original set of reference sites. We compared the number of outliers identified based on the revised set of reference sites, with the number of outliers identified based on the original set of reference sites. To further test whether the revised models assessed a larger range of naturally occurring site conditions than the original models, we calculated the % increase in the range of values for the same 5 environmental predictor variables used to assess reference-condition environmental space.

### Using prediction intervals to set site-specific benchmarks

We compared 2 approaches to derive PIs for each of the revised random forest models. First, we derived PIs from quantile regression forests (QRFs). QRFs are a generalization of random forests that can be used to empirically derive prediction intervals (Hengl et al. 2018). When used for regression, random forest models provide an accurate approximation of the conditional mean of a response variable, given specific values of the predictor variables. QRFs, on the other hand, estimate conditional quantiles, not just the mean. We used QRF to predict the 95th percentiles of the SC, [TN], and [TP] values at test sites. We implemented QRF with the *quantregForest* package (version 1.3-7; Meinshausen 2024) in R.

In addition to the QRF method, we derived PLs from the simple empirical error (SEE) method (Olson and Hawkins 2013). Briefly, the SEE method involves bootstrapping the residuals from the reference data and adding each bootstrapped residual to the prediction to create an empirical distribution of the prediction + error. For each test-site prediction, we sampled the residuals 500× with replacement and added each sampled residual to the prediction to produce a distribution of the prediction + error. We then chose the 95th percentile of that distribution as the upper PL for that prediction. We used boxplots to visually compare upper PLs for each water-quality indicator and used paired Wilcoxon rank-sum tests to assess whether upper PLs differed between the QRF and SEE methods. We also compared these 2 approaches by calculating the number of test sites that exceeded the 95th percentile values for each approach. We used the 95th percentile as a threshold to illustrate differences between the 2 approaches, but we are not necessarily advocating its use in a formal monitoring or regulatory context.

### Comparing site-specific and regional benchmarks

We compared site-specific benchmarks (i.e., 95th percentiles) derived from the QRF and SEE methods with regional benchmarks used by the EPA's NRSA. First, we allocated test sites to 1 of the 4 Omernik level III ecoregions in the study area. Next, we calculated the percentage of sites within each ecoregion that exceeded regional NRSA, SEE-derived, and QRF-derived benchmarks. Regional NRSA benchmarks for SC, [TN], and [TP] were obtained from USEPA (2023). Regional benchmarks for SC were set to either 1000 or 2000 μS/cm, depending on the ecoregion, and were largely based on best professional judgment (USEPA 2023). Regional benchmarks for [TN] and [TP] were derived by the EPA as the 95th percentile of the reference distributions for each ecoregion. Regional benchmarks for [TN] ranged from 249 to 1069 μg/L, and regional benchmarks for [TP] ranged from 41 to 127 μg/L (USEPA 2023).

## RESULTS

### Reference and test sites

Overall, test-site watersheds contained lower percentages of forested land cover and higher percentages of urban and agricultural land cover than reference-site watersheds (Dewitz 2023) (Fig. S2). On average, SC test-site watersheds contained 38.2% (range 0–100%) forested land cover, whereas the original and revised SC reference-site watersheds contained 62.3% (range 0–100%) and 56.2% (range 0–100%) forested land cover on average, respectively. The mean % cover of urban and agriculture land use in SC test watersheds was 0.1% (range 0–9.1%) and 0.7% (range 0–59.7%), respectively. On average, SC reference-site watersheds contained <1% urban and agricultural land cover. For TN and TP, test-site watersheds contained 38.9% (range 0–98.2%) and 39.0% (range 0–98.2%) forested land cover on average, respectively. TN and TP reference-site watersheds had >60% forested land cover on average (range for TN and TP 0–100%). Agricultural land cover at TN and TP test sites ranged from 0% to 60%, whereas agricultural land cover at TN and TP reference sites ranged from 0% to 10% (Fig. S2). On average, urban land cover constituted <1% of TN and TP reference-site and test watersheds.

### Predictor variables

Ten predictor variables were included in the revised SC model, whereas 19 predictor variables were included in the original SC model (Table 1). Predictors in both the revised and original SC models mainly represented climatic and lithological associations with SC. The difference between precipitation and evapotranspiration (Precip_Minus_EVTWs) and the percentages of $Al_2O_3$, CaO, and S were most strongly associated with SC in the revised model (Table 1). Precip_Minus_EVTWs and % $Al_2O_3$ in the surface lithology were negatively related to SC, whereas % CaO and % S were positively related to SC (Table 1, Fig. 2A). Both maximum and minimum temperatures (Tmax8110Ws and Tmin8110Ws; 30-y normals for 1981–2010) were strong predictors and positively related to SC in the revised model (Table 1, Fig. 2A). In comparison, % CaO, % S, maximum temperature, mean number of wet days, and mean annual precipitation were the predictors most strongly associated with SC in the original SC model (Table 1; see Table S2 for descriptions of the predictors used in the original models). Similar to the revised model, predictors related to precipitation were negatively associated with SC, whereas temperature-related predictors were positively associated with SC (Table 1). Overall, the revised and original models had 3 predictor variables in common (% CaO, % S, and maximum temperature).

Seven predictors were included in the revised TN model, whereas the original TN model included 12 predictor variables

Table 1. List of the predictors included in the revised and original predictive models of spatial variation in specific conductivity (SC), total N (TN), and total P (TP) at reference sites in the western United States. Var imp is the variable importance calculated as the % increase in mean square error when the predictor is randomly permuted. Direction indicates the direction of the overall association between the predictor and the water-chemistry constituent. Predictors that are common between the original and revised models are in italics and have the same superscripts. Note that names for the same predictor variables can differ between the original and revised models. Predictors that original and revised models have in common: CaOWs = % CaO, SWs = % S, Tmax8110Ws = maximum temperature, DOY = day of the year. Ws refers to watershed-scale predictors from the StreamCat dataset. Detailed descriptions of predictor variables can be found in Table S1 (revised models) and Table S2 (original models). N/A = not applicable.

| | Revised models | | | | Original models | | | |
|---|---|---|---|---|---|---|---|---|
| Model | Predictor name | Units | Var imp | Direction | Predictor name | Units | Var imp | Direction |
| SC | Precip_Minus_EVTWs | km/km$^2$ | 47 | − | *% CaO*[1] | % | 63 | + |
| | Al2O3Ws | % | 40 | − | *% S*[2] | % | 42 | + |
| | *CaOWs*[1] | % | 30 | + | *Maximum temperature*[3] | °C | 41 | + |
| | *SWs*[2] | % | 26 | + | Mean wet days | d/y | 37 | − |
| | Tmin8110Ws | °C | 25 | + | Mean precipitation | mm/y | 35 | − |
| | NWs | % | 25 | + | Soil bulk density | g/cm$^3$ | 33 | + |
| | *Tmax8110Ws*[3] | °C | 24 | + | Soil permeability | inches/h | 33 | − |
| | ElevWs | m | 24 | − | Atmospheric Mg | mg/L | 32 | + |
| | RunoffWs | mm | 22 | − | Atmospheric Ca | mg/L | 32 | + |
| | BFIWs | ratio | 17 | − | % MgO | % | 32 | + |
| | | | | | Atmospheric $SO_4$ | mg/L | 31 | + |
| | | | | | Mean maximum EVI | N/A | 30 | + |
| | | | | | Compressive strength | MPa | 30 | − |
| | | | | | Minimum precipitation | mm/mo | 29 | − |
| | | | | | Max wet days | d/y | 28 | − |
| | | | | | Soil erodibility | N/A | 28 | + |
| | | | | | Day last freeze | Day of the year | 28 | − |
| | | | | | Log hydraulic conductivity | log(m/s) | 27 | + |
| | | | | | Mean summer precipitation | mm/mo | 24 | − |
| TN | Precip8110Ws | mm/y | 22 | − | Mean wet days | d/y | 27 | − |
| | Precip_Minus_EVTWs | km/km$^2$ | 20 | − | Minimum temperature | °C | 25 | + |
| | ElevWs | m | 19 | − | Atmospheric Na | mg/L | 24 | + |
| | RunoffWs | mm | 16 | − | *Day of the year*[4] | d | 24 | − |
| | BFIWs | ratio | 16 | − | Prior 2-mo precipitation | mm/mo | 23 | − |
| | PermWs | cm/h | 12 | − | Atmospheric $NO_3$ | mg/L | 21 | − |
| | *DOY*[4] | Day of the year | 10 | − | | | | |
| | | | | | Atmospheric $SO_4$ | mg/L | 21 | + |
| | | | | | EVI | N/A | 20 | + |
| | | | | | Soil bulk density | g/cm$^3$ | 18 | − |
| | | | | | Ground water index | N/A | 16 | − |
| | | | | | % evergreen | % | 15 | − |
| | | | | | % *Alnus rubra* dominated | % | 10 | + |
| TP | Precip8110Ws | mm/y | 26 | − | Gila Mountains ecoregion | N/A | 36 | + |
| | Precip_Minus_EVTWs | km/km$^2$ | 24 | − | % volcanic lithology | % | 31 | + |

Table 1. (*Continued*)

| | Revised models | | | | Original models | | | |
|---|---|---|---|---|---|---|---|---|
| Model | Predictor name | Units | Var imp | Direction | Predictor name | Units | Var imp | Direction |
| | P2O5Ws | % | 22 | + | Previous year's precipitation | mm/y | 26 | − |
| | ClayWs | % | 21 | + | *% CaO*[1] | % | 24 | − |
| | *CaOWs*[1] | % | 20 | − | Relative humidity | % | 24 | − |
| | Tmax8110Ws | °C | 20 | + | Minimum temperature | °C | 22 | + |
| | RunoffWs | mm | 19 | − | Area largest water body | $m^2$ | 21 | − |
| | SandWs | % | 17 | − | Mean channel slope | % | 21 | − |
| | WetIndexWs | N/A | 15 | + | Atmospheric Ca | mg/L | 21 | + |
| | | | | | SOC | kg C/$m^2$ | 20 | − |
| | | | | | EVI | N/A | 19 | + |
| | | | | | Soil water capacity | fraction | 19 | + |
| | | | | | Soil erodibility (K factor) | N/A | 18 | − |
| | | | | | % P | % | 16 | + |
| | | | | | % Alfisols | % | 15 | + |

(Table 1). Precipitation-related predictors (Precip8110Ws and Precip_ Minus_EVTWs), elevation (ElevWs), and predictors related to water flow (RunoffWs, BFIWs) were most strongly associated with [TN] in the revised model (Table 1). Precip8110Ws and Precip_Minus_EVTWs were negatively associated with [TN] (Fig. 2B). ElevWs, RunoffWs, BFIWs, and the mean permeability of soils within a watershed (PermWs) were also important predictors (Table 1) and were negatively related to [TN] (Fig. 2B) in the revised TN model. In the original TN model, the mean number of wet days, minimum temperature, and atmospheric Na deposition were most strongly associated with [TN] (Table 1). DOY was included in both the revised and original TN models, but it was relatively more important in the original model than in the revised model (% increase in MSE of 24% in the original model vs 10% in the revised model; Table 1).

The revised TP model included 9 predictor variables, whereas the original TP model included 15 predictors (Table 1). In the revised model, Precip8110Ws, Precip_ Minus_ EVTWs, % $P_2O_5$, and % clay content of the soils (ClayWs) were most strongly associated with [TP] (Table 1). Similar to [TN], Precip8110Ws and Precip_Minus_EVTWs were negatively related to [TP] (Table 1, Fig. 2C). The percentages of $P_2O_5$, clay, and CaO in the surface lithology were also strong predictors of [TP] (Table 1). Both P2O5Ws and ClayWs were positively related to [TP], whereas CaOWs was negatively related to [TP] (Table 1, Fig. 2C). The most important predictors of [TP] in the original model were % volcanic lithology, the previous year's annual precipitation, % CaO, and whether or not a site was located within the Gila Mountains ecoregion, which contains large amounts of young basalt rocks (Table 1). The % CaO content of the soil was the only predictor the revised and original models had in common.

## Model development and evaluation

The revised and original SC models were similar in terms of performance. The revised SC model explained ~74% of the variation in SC across reference sites, whereas the original model explained 78% (Table 2), and both were unbiased estimators of SC (Table S3, Fig. 3A). The RMSE of the revised SC model was ~8% of the range of observed SC (range 4–979 μS/cm), and the MAE between predicted and observed SC was 50.5 μS/cm. The original SC model had an RMSE that was ~6.5% of the range of observed SC values (range 133–1171 μS/cm). The MAE of the original model was 42.2 μS/cm.

Predicted SC levels by the revised SC model were slightly higher, on average, than those made by the original SC model for reference sites but slightly lower, on average, for test sites (Table 3). On average, the revised model predicted reference-site SC values that were 3.3 μS/cm higher than those predicted by the original SC model (ranges 12.6–836.9 μS/cm and 11.5–825.1 μS/cm for revised and original models, respectively). Predictions of SC at reference sites by the original model were highly correlated with predictions by the revised model ($r$ = 0.94), and the slope of the relationship between the 2 predictions was statistically indistinguishable from 1 (slope = 1.00, 95% CI = 0.98–1.02; Table S3, Fig. 4A). Revised model predictions of naturally occurring SC levels at test sites (range 30.6–685.6 μS/cm) were, on average, 2.7 μS/cm lower than predictions by the original

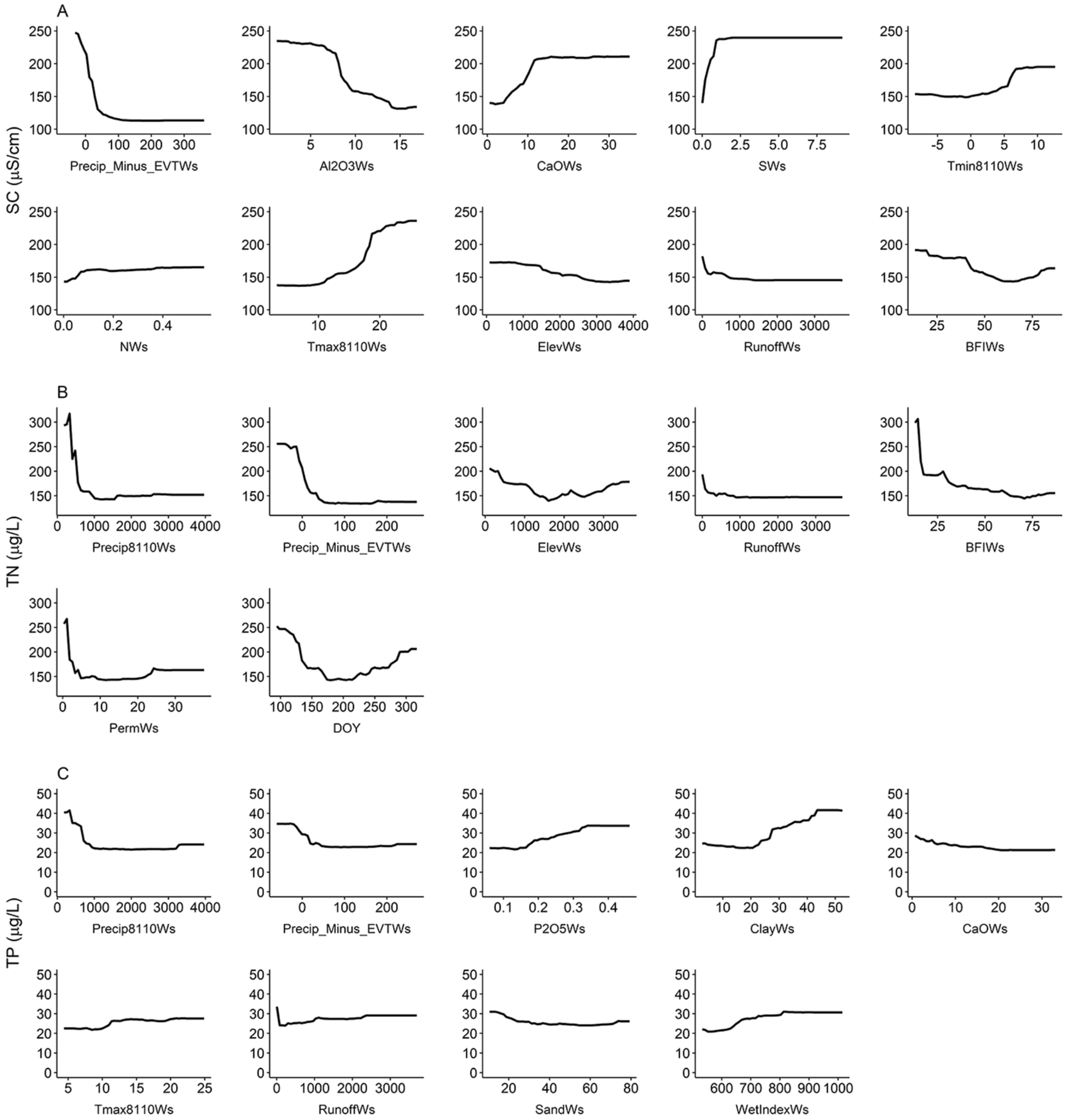

Figure 2. Partial dependence of predictor variables for the revised specific conductivity (SC; $n = 1912$) (A), total N (TN; $n = 699$) (B), and total P (TP; $n = 966$) (C) random forest models at reference sites in the western United States. Precip_Minus_EVTWs = precipitation minus evapotranspiration in the watershed; AL2O3Ws = watershed mean % lithological $Al_2O_3$ content in surface or near-surface geology; CaOWs = watershed mean % lithological CaO content in surface or near-surface geology; SWs = watershed mean % lithological S content in surface or near-surface geology; Tmin8110Ws = 30-y normal min. temperature (°C) in the watershed; NWs = watershed mean % lithological N content in surface or near-surface geology; Tmax8110Ws = 30-y normal max. temperature (°C) in the watershed; ElevWs = mean watershed elevation (m); RunoffWs = mean runoff (mm) in the watershed; BFIWs = baseflow index in the watershed; Precip8110Ws = 30-y normal mean precipitation in the watershed; PermWs = mean watershed soil permeability (cm/h); DOY = day of the year; P2O5Ws = watershed mean % of lithological $P_2O_5$ content in surface or near-surface geology; ClayWs = mean % clay content of soils in the watershed; SandWs = mean % sand content of soils; WetIndexWs = mean composite topographic index in the watershed.

model (range 23.0–649.0 µS/cm; Table 3, Fig. 4B). Revised and original model predictions were also highly correlated at test sites ($r = 0.91$). However, the slope of the relationship between original and revised predictions was >1 (slope = 1.11, 95% CI = 1.09–1.14), and the intercept was <0 (intercept = −24.32, 95% CI = −30.29 to −18.49; Table S3, Fig. 4B), indicating there was systematic bias in the predictions by one or both models.

The revised TN model explained slightly more of the variation in [TN] across reference sites ($r^2 = 0.39$) (Fig. 3B) than the original TN model ($r^2 = 0.32$; Table 2). The RMSE of the revised TN model was ~14% of the range of observed

Table 2. Metrics of model performance for the original and revised water-quality models of spatial variation in specific conductivity (SC), total N (TN), and total P (TP) at reference sites in the western United States. Metrics are based on the relationship between the observed and internal out-of-bag predicted values for reference sites. NSE is the Nash–Sutcliffe model efficiency, RMSE is the root mean square error, and MAE refers to the mean absolute error.

| Model | Indicator | *n* | $r^2$ | NSE | RMSE | MAE | No. predictors |
|---|---|---|---|---|---|---|---|
| Original models | SC | 1390 | 0.78 | 0.78 | 67.3 | 42.2 | 19 |
| | TN | 665 | 0.32 | 0.32 | 113.9 | 71.0 | 9 |
| | TP | 752 | 0.40 | 0.40 | 20.5 | 10.6 | 15 |
| Revised models | SC | 1912 | 0.74 | 0.74 | 81.1 | 50.5 | 10 |
| | TN | 699 | 0.39 | 0.39 | 120.3 | 77.7 | 7 |
| | TP | 966 | 0.38 | 0.38 | 19.4 | 13.1 | 9 |

values (range 5–879 μg/L), whereas it was 11.9% of the range of observed values (range 5–960 μg/L) in the original model. For the revised TN model, MAE was 77.7 μg/L, whereas it was 71.0 μg/L for the original TN model.

Model predictions of [TN] were higher for revised than original models at both reference and test sites. Revised model predictions of [TN] at reference sites were slightly higher (2.8 μg/L), on average, than predictions by the original model (ranges 45.3–534.0 μg/L and 33.5–528.4 μg/L for revised and original models, respectively; Table 3). Predictions by the 2 models were moderately and positively correlated ($r = 0.79$), but the models were slightly biased predictors of one another (slope = 1.08, 95% CI = 1.02–1.15; Table S3, Fig. 4C). For test sites, revised model predictions were, on average, 33.3 μg/L higher than predictions by the original model (ranges 58.6–742.7 and 49.5–402.0 μg/L for revised and original models, respectively; Table 3). Predictions by the revised and original models at test sites were less strongly correlated ($r = 0.51$) with one another than those observed for reference sites, but the slope of the relationship between the 2 models' predictions approximated 1 (slope = 0.92, 95% CI = 0.84–1.01; Table S3, Fig. 4D).

The revised TP model performed similarly to the original TP model in terms of $r^2$ (0.38 vs 0.40), NSE (0.38 vs 0.40), RMSE (19.4 vs 20.5 μg/L), and MAE (13.1 vs 10.6 μg/L) (Table 2). Predicted values at reference sites ranged between 1.3 and 92.6 μg/L for the revised model (Fig. 3C) and between −0.7 and 77.6 μg/L for the original model (Table 3). Predictions by the 2 models were moderately correlated ($r = 0.78$), and the slope of their relationship was not statistically different from 1 (slope = 1.06, 95% CI = 1.00–1.12; Table S3, Fig. 4E). For test sites, the revised model predictions ranged from 6.0 to 85.6 μg/L, whereas predictions by the original model ranged from 4.7 to 63.5 μg/L (Table 3). Predictions by the 2 models at test sites were only weakly ($r = 0.58$) correlated with one another, and the revised and original models were biased predictors of one another (slope = 0.73, 95% CI = 0.68–0.79; Table S3, Fig. 4F).

## Identifying sites outside reference-site environmental space

The environmental space covered by the revised models slightly increased with the addition of new reference sites.

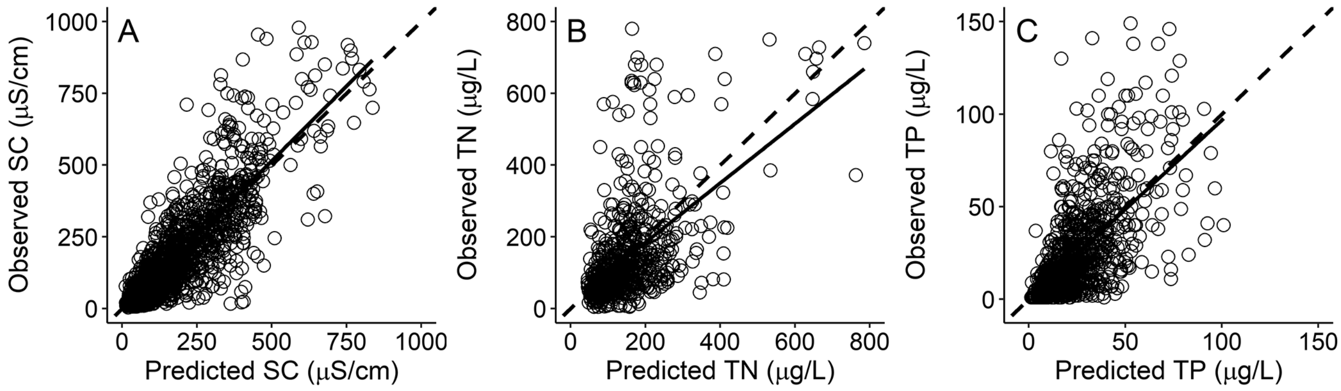

Figure 3. Plots of observed values as a function of predicted values from revised models at reference sites in the western United States for specific conductivity (SC; $n = 1912$) (A), total N (TN; $n = 699$) (B), and total P (TP; $n = 966$) (C). The dashed line is the 1:1 line, and the solid line is the regression line based on ordinary least squares linear regression. Reference-site data for SC ranged from 1965 to 2016, and data for TN and TP ranged from 1973 to 2015 and 1973 to 2019, respectively.

Table 3. Summary statistics for predictions made using the original and revised water-quality models for reference and Bureau of Land Management's Assessment, Inventory and Monitoring Strategy test sites in the western United States. Difference was calculated as the original predictions subtracted from the revised predictions for each site.

| | | Original models | | | Revised models | | | Difference | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Site type | Indicator | Mean | Min. | Max. | Mean | Min. | Max. | Mean | Min. | Max. |
| Reference | SC | 132.6 | 11.5 | 825.1 | 135.9 | 12.6 | 836.9 | 3.3 | −363.8 | 363.0 |
| | TN | 142.0 | 33.5 | 528.4 | 144.8 | 45.3 | 534.0 | 2.8 | −213.8 | 299.9 |
| | TP | 20.0 | −0.7 | 77.6 | 22.0 | 1.3 | 92.6 | 2.0 | −28.5 | 65.1 |
| Test | SC | 238.9 | 23.0 | 649.0 | 236.2 | 30.6 | 685.6 | −2.7 | −320.2 | 335.1 |
| | TN | 191.7 | 49.5 | 402.0 | 225.1 | 58.6 | 742.7 | 33.4 | −207.4 | 485.2 |
| | TP | 28.1 | 4.7 | 63.5 | 37.5 | 6.0 | 85.6 | 9.4 | −34.0 | 55.1 |

For example, the revised SC and TP models were more applicable to sites with smaller watersheds than the original models (Table 4). The original SC and TP models were trained on sites with watersheds that ranged in size from 0.6 to 25,362 km$^2$. The revised SC model was trained on sites with watersheds as small as 0.02 km$^2$, and the revised TN model was applicable to watersheds as small as 0.4 km$^2$. We were unable to add reference sites with watersheds >25,362 km$^2$. In addition, the revised SC model was applicable to sites at slightly lower and higher elevations than the original SC model (Table 4). The revised SC model was trained on sites that ranged from 82 to 3897 m a.s.l., whereas the original SC model was trained on sites that ranged from 109 to 3773 m a.s.l. (an increase in range of 4.1%). The ranges in minimum and maximum temperatures were also slightly increased for the revised SC model (increases of 4.0% and 3.2%, respectively; Table 4). Adding additional reference sites did not increase the range of precipitation covered for any of the models.

Overall, slightly fewer test sites were assessed as being outside reference-site environmental space as defined by the revised set of reference sites than by the original reference sites. Of the 1957 test sites with SC values, 458 (23.4%) fell outside the environmental space of the revised model's reference sites vs 481 (24.6%) for the original model. For TN, 290 out of 1539 (18.8%) test sites were assessed as being outside reference-site environmental space by the revised model reference sites vs 314 (20.4%) for the original model reference sites. Of the 1547 test sites with [TP] values, 257 (16.6%) test sites were assessed as being outside the revised model's reference-site environmental space vs 354 (22.9%) for the original model's reference sites. In general, sites that were flagged as being outside the reference-site environmental space had larger watersheds, higher temperatures, lower amounts of precipitation, and were at lower elevations (Fig. S3). Based on the results of these tests, we could confidently assess the predicted results from the revised models for 1499 (~77%) SC test sites, 1249 (~81%) TN test sites, and 1280 (~83%) TP test sites (Fig. S4).

### Using prediction intervals to set site-specific benchmarks

In general, QRF produced higher site-specific upper PLs (95$^{th}$ percentile values) than the SEE method. For SC, upper PLs derived from QRF ranged from 70.1 to 940.0 μS/cm, whereas upper PLs from the SEE method ranged from 141.0 to 808.4 μS/cm (Fig. 5A). On average, upper PLs for SC derived from QRF were ~100 μS/cm higher than those derived from the SEE method (Wilcoxon rank-sum test $W = 825{,}410$, $p < 0.001$). Of the 1499 SC test sites, the QRF method flagged 21.4% as having SC values greater than the upper PL, whereas the SEE method flagged 31.6% of test sites (Table 5). Upper TN PLs produced by QRF ranged from 88.0 to 879.0 μg/L, and those produced by the SEE method ranged from 245.8 to 1000.8 μg/L. Similar to upper PLs for SC, QRF upper PLs for TN were higher than those produced by SEE (Wilcoxon rank-sum test $W = 474{,}186$, $p < 0.001$; Fig. 5B). On average, QRF upper PLs were 117 μg/L higher than those derived from SEE. Of the 1249 TN test sites, QRF PLs flagged ~19.1% of sites and SEE PLs flagged 25.6% of sites as having excess [TN] (Table 5). Site-specific upper TP PLs derived from QRF ranged from 13.1 to 149.0 μg/L. Upper TP PLs produced by the SEE method ranged from 38.6 to 125.7 μg/L and were generally 10 μg/L lower than upper PLs produced by QRF (Wilcoxon rank-sum test $W = 625{,}897$, $p < 0.001$; Fig. 5C). Of the 1280 TP test sites, 23.2% had observed [TP] greater than the upper PL produced by QRF, whereas 26.2% had observed [TP] greater than the upper PL produced by SEE (Table 5).

### Comparing site-specific and regional benchmarks

For each water-chemistry constituent, the percentage of test sites that exceeded benchmarks varied depending on whether regional or site-specific benchmarks were used. Site-specific SC benchmarks (SEE and QRF) flagged a higher percentage of test sites as being in nonreference condition than regional-derived benchmarks. Only 9.1%

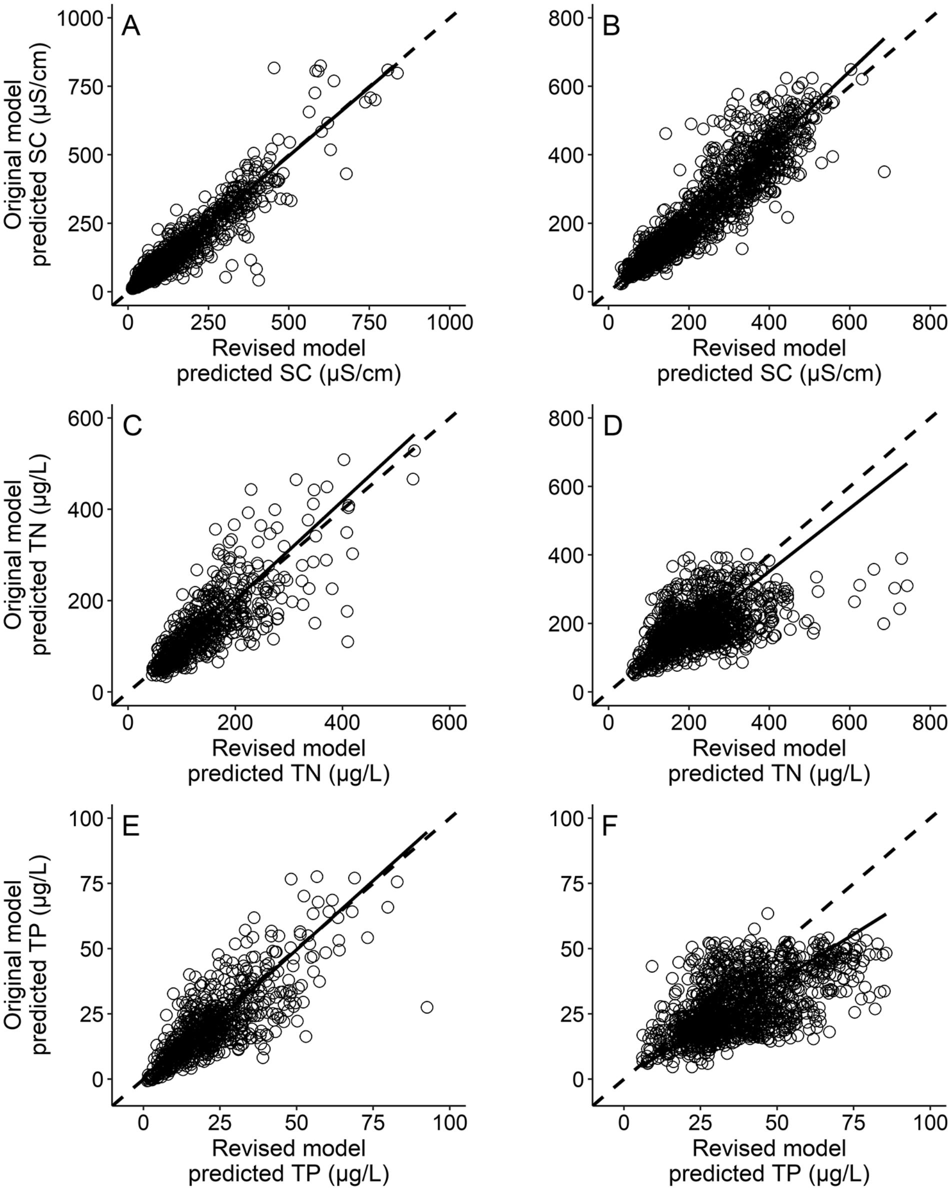

Figure 4. Scatterplots depicting predicted water-quality indicator values at reference (A, C, E) and test (B, D, F) sites in the western United States derived from the original models (*y*-axis) vs the revised models (*x*-axis). Predictions are for specific conductivity (SC) (A, B), total N (TN) (C, D), and total P (TP) (E, F). The dashed line is the 1:1 line and the solid line is the regression line based on reduced major axis (RMA) regression. RMA statistics can be found in Table S2. Reference-site data for SC ranged from 1965 to 2016, and data for TN and TP ranged from 1973 to 2015 and 1973 to 2019, respectively. Test-site data ranged from 2013 to 2021.

of SC test sites exceeded NRSA regional benchmarks, whereas a much higher percentage of sites (21.4–31.6%) exceeded site-specific SC benchmarks (Table 5). In contrast, for both [TN] and [TP], the use of regional NRSA benchmarks flagged a higher percentage of sites with excessive concentrations than site-specific benchmarks did. Specifically for [TN], the use of regional NRSA benchmarks flagged 31.1% of test sites, whereas QRF and SEE benchmarks flagged 19.1% and 25.6%, respectively (Table 5). For [TP], regional benchmarks flagged 33.2% of test sites as being in nonreference condition, but only 23.2 to 26.3% of test sites were flagged when assessed against site-specific benchmarks (Table 5). The percentage of test sites flagged as being in nonreference condition also varied among ecoregions for each water-chemistry constituent (Table S4).

## DISCUSSION

### Revised and original model performance

Our primary goal was to improve the performance of existing empirical models used to predict naturally occurring spatial variation in SC, [TN], and [TP]. We also wanted to

Table 4. The minimum and maximum values of 5 environmental variables used to assess the range in natural environmental conditions of reference sites in the western United States for original and revised random forest models predicting spatial variation in water-quality indicators: specific conductivity (SC), total N (TN), and total P (TP). WsAreaSqKm = watershed area (km$^2$); Tmax8110Ws = mean maximum temperature from 1981 to 2010 (°C) within the watershed; Tmin8110Ws = mean minimum temperature from 1981 to 2010 (°C) within the watershed; Precip8110Ws = mean annual precipitation from 1981 to 2010 (mm) within the watershed; ElevWs = mean watershed elevation (m). % increase is the percentage of increase in the range of values.

| | | Original | | Revised | | |
|---|---|---|---|---|---|---|
| Indicator | Predictor | Min. | Max. | Min. | Max. | % increase |
| SC | WsAreaSqKm | 0.6 | 25362.4 | 0.02 | 2707.8 | <1 |
| | Tmax8110Ws | 3.9 | 26.1 | 3.2 | 25.7 | 3.2 |
| | Tmin8110Ws | −8.2 | 12.0 | −8.4 | 12.6 | 4.0 |
| | Precip8110Ws | 179.8 | 5195.4 | 193.3 | 3998.8 | 0 |
| | ElevWs | 108.8 | 3773.4 | 82.0 | 3897.0 | 4.1 |
| TN | WsAreaSqKm | 0.6 | 25362.4 | 0.4 | 6098.3 | <1 |
| | Tmax8110 | 3.9 | 26.1 | 5.9 | 21.4 | 0 |
| | Tmin8110Ws | −8.2 | 12.0 | −7.0 | 7.5 | 0 |
| | Precip8110Ws | 179.8 | 5195.4 | 182.6 | 3599.5 | 0 |
| | ElevWs | 108.8 | 3773.4 | 232.0 | 3260.0 | 0 |
| TP | WsAreaSqKm | 0.6 | 25362.4 | 1.5 | 6098.2 | 0 |
| | Tmax8110Ws | 3.9 | 26.1 | 5.5 | 15.3 | 0 |
| | Tmin8110Ws | −8.2 | 12.0 | −5.7 | 0.9 | 0 |
| | Precip8110Ws | 179.8 | 5195.4 | 305.3 | 1286.0 | 0 |
| | ElevWs | 108.8 | 3773.4 | 755.8 | 2798.4 | 0 |

create models that were easier to compute and implement, applicable to a wider range of environmental settings, free from shifting-baseline issues, and able to provide site-specific benchmarks. Although we did not achieve marked improvement in model accuracy and precision, the 3 revised models do provide the following advantages over the original models: 1) the revised models use reproducible predictors from a consistent, standard, nationally available dataset (StreamCat), which reduces the need for advanced GIS expertise; 2) the revised models were trained on a larger number of reference sites, which slightly increased the environmental space to which the models can be applied; and 3) the revised models are less susceptible to shifting-baseline issues than other predictive models (e.g., Olson and Cormier 2019).

All 3 revised models produced relatively accurate predictions, but model precision for [TN] and [TP] was still relatively poor. Olson and Hawkins (2013) attributed the poor precision of the original models to temporal and measurement variation in SC, [TN], and [TP], as well as changes in analytical methods by sampling agencies over time. These same issues likely contributed to the poor precision of the revised models because the original and revised models share many of the same reference sites. Some of the unexplained variation in SC, [TN], and [TP] is also likely associated with variation in reference-site quality (reference sites are minimally disturbed and most are not pristine). Reference sites vary in their quality (i.e., how much land use has occurred in their watersheds). For example, the amount of agriculture within our population of reference-site watersheds varied from 0% to 8% (Fig. S2). In addition, neither the predictors used nor the spatial and temporal resolution at which they were measured likely captured important naturally occurring biogeochemical processes, especially those influencing [TN], and to a lesser extent [TP], dynamics. For example, the TN model included 1 relatively weak soil predictor (permeability), which may be associated with N fixation and subsequent leaching to streams. However, we did not identify any other predictor associated with N fixation, denitrification, or nutrient uptake, which can strongly influence background nutrient concentrations in streams (Webster et al. 2016). Furthermore, it was striking that several of the selected predictors for the TN model characterized abiotic processes associated with dilution or mobilization (e.g., precipitation, runoff, baseflow), not sources of N (e.g., % rock N). Accounting for local biogeochemical processes in future models should increase our ability to predict nutrient concentrations in freshwater ecosystems, but doing so remains a challenge because some of these processes are still poorly understood at large scales (Marcarelli et al. 2022).

In some cases, predictors exist that could potentially better capture effects of local biogeochemical processes,

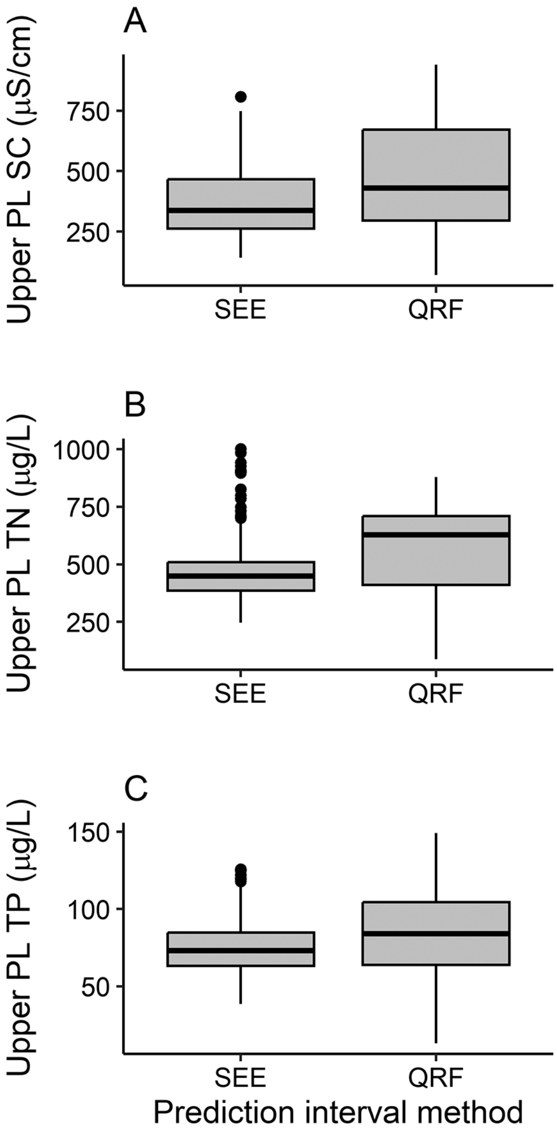

Figure 5. Boxplots depicting the range in upper prediction limits (upper PL) estimated from the simple empirical error (SEE) method and quantile regression forests (QRF) for western United States test-site predictions made with the revised specific conductivity (SC) (A), total N (TN) (B), and total P (TP) (C) models. The lower and upper hinges correspond to the 25th and 75th percentiles, respectively. The lower and upper whiskers extend no further than 1.5× the interquartile range. Points represent outliers.

but we did not include them in our models because they can be altered by anthropogenic activity. For example, land cover and atmospheric N deposition predictors were readily available and have been shown to be associated with spatial variation in nutrient concentrations (Lek et al. 1999, Greathouse et al. 2014, Shen et al. 2020, Lin et al. 2021). However, it is difficult to determine how much spatial variation in land cover or atmospheric N deposition is natural and how much has been altered by human activities (e.g., logging, burning of fossil fuels).

Some researchers have used both natural predictors and predictors altered by human activity (e.g., urban and agricultural land cover) in models (Smith et al. 2003, Dodds and Oakes 2004). These models often account for more spatial variation across sites than our revised models, but they are typically used to understand the relative importance of anthropogenic sources of ions to streams rather than to predict reference-site concentrations. These models could be used to extrapolate to the concentrations expected under reference conditions by setting land-use disturbance to 0, but such extrapolations suffer from much higher uncertainty than the type of interpolations made by reference-quality models. Additionally, the relationships between ion concentrations and predictors such as land use are often nonlinear, making linear models inappropriate for such predictions (Dodds et al. 2010).

### Benefits and drawbacks of using a standardized, nationally available dataset of predictor variables

Some predictive models can also be difficult to implement. The original water-quality models of Olson and Hawkins (2012, 2013) and Le et al. (2019) incorporated complicated predictor variables that require watershed delineations and special GIS layers from multiple sources (e.g., LANDFIRE, MODIS). Moreover, several of the predictor variables were calculated with GIS software that requires expertise in geospatial analysis. For example, to characterize the amount of rock–water contact occurring in a site's watershed, the original TN model uses an index of groundwater velocity as one of the input variables. This index is estimated using a magnetic resonance imaging model (Baker et al. 2003) applied within a GIS environment, which can be a complicated procedure for nonexpert GIS practitioners. Our goal was to create models that would be easy to compute, understand, and implement for all end users. With the development of the StreamCat dataset (Hill et al. 2016), there is less need for geospatial expertise to calculate landscape metrics. Our revised models incorporate only StreamCat variables, allowing end users to quickly and easily extract predictors.

Our predictive models based on the StreamCat dataset allow users to easily make predictions of water chemistry for ~2.1 million river km and their associated catchments in the western United States. Importantly, these predictions can be made without first having to delineate watersheds and calculate predictor variables within a GIS environment. Spatially explicit maps of such fine-scale predictions could be used by scientists and land managers to visualize spatial patterns in water chemistry and rapidly identify stream reaches at risk of impairment or reaches that should be prioritized for protection. Although it was beyond the scope of this project, predictions of reference-condition water chemistry could be

Table 5. The number of western United States test sites at which observed values exceeded the 95th percentile vs the number of test sites at which observed values were below the 95th percentile for each benchmark method. The percentages of sites within each category are in parentheses. QRF = quantile random forest, SEE = simple empirical error method, NRSA = National Rivers and Streams Assessment. Test-site data were obtained from the Bureau of Land Management's Assessment, Inventory and Monitoring Strategy Lotic Indicators Hub (https://gbp-blm-egis.hub.arcgis.com/pages/aim, accessed August 2023) for models of 3 water-quality indicators: specific conductivity (SC), total N (TN), and total P (TP).

| | | QRF | | SEE | | NRSA | |
|---|---|---|---|---|---|---|---|
| Indicator | *n* | Exceeded 95th percentile (no.) (% of sites) | Below 95th percentile (no.) (% of sites) | Exceeded 95th percentile (no.) (% of sites) | Below 95th percentile (no.) (% of sites) | Exceeded 95th percentile (no.) (% of sites) | Below 95th percentile (no.) (% of sites) |
| SC | 1499 | 321 (21.4) | 1178 (78.6) | 474 (31.6) | 1025 (68.4) | 136 (9.1) | 1363 (90.9) |
| TN | 1249 | 238 (19.1) | 1011 (80.9) | 320 (25.6) | 929 (74.4) | 388 (31.1) | 861 (68.9) |
| TP | 1280 | 297 (23.2) | 983 (76.8) | 337 (26.3) | 943 (73.7) | 425 (33.2) | 855 (66.8) |

expanded to over 4.2 million river km in the conterminous United States if the models were trained on an expanded reference dataset.

The ease of obtaining predictor values without having to first delineate watersheds comes at a cost, however. For example, metrics in the StreamCat dataset characterize the drainage area contributing to an entire NHD reach, which can be relatively large, rather than the specific sample point. This imprecision in specifying a site's location and its associated watershed could contribute to imprecise or inaccurate predictions, especially for small headwater systems that do not fall on the NHD network. However, despite the imprecise matching of site location and associated StreamCat predictors, we were still able to produce models of comparable performance to the original models that were based on point-specific metrics.

The StreamCat dataset does not include all possible variables that could influence streamwater chemistry, but it provides a useful set of predictors for modeling water quality. Ideally, the predictor variables used in these models should be interpretable in terms of the mechanisms known to influence the water-chemistry constituent being predicted. The StreamCat dataset contains several predictors that can be mechanistically linked to SC, [TN], and [TP] in surface waters, although it is missing some potentially important predictors (e.g., cover of N-fixing plants in the watershed). The StreamCat dataset contains predictor variables related to 3 general processes that control concentrations of ions in surface waters: 1) atmospheric precipitation, 2) the mineral composition and weathering of rock, and 3) evaporation (Gibbs 1970). Variables associated with precipitation, evapotranspiration, and lithology were consistently the most important predictors in the 3 revised models. Most, if not all, of the relationships between predictors and the water-quality constituents were interpretable (e.g., higher precipitation associated with lower [TN] and [TP] because of dilution). Thus, the StreamCat and similar regional and global datasets (e.g., HydroSHEDS, PRISM) can produce interpretable predictive models that perform well enough to address many water-resource-management objectives.

### Model applicability

Another goal in revising the original models was to make the models applicable to a larger naturally occurring environmental space. We trained all 3 revised models on a greater number of reference sites than the original models, but doing so did not appreciably increase the range of natural site conditions covered by the models. In particular, we were unable to make the models applicable to sites with watersheds larger than 25,362 km$^2$, which was one of our goals. However, we did add reference sites from smaller watersheds (<0.6 km$^2$) for the SC and TP models, which were previously underrepresented. We were also able to increase the elevation range slightly for the SC model. Thus, the revised SC and TP models should be applicable to more headwater mountain streams than the original models.

The addition of new reference sites also resulted in fewer test sites being identified as outside of reference-site environmental space for all 3 models, despite the small increases in the range of naturally occurring reference-site conditions. This result suggests that we indeed increased the coverage of multidimensional environmental space by including the new reference sites. Natural resource managers should therefore be able to assess a larger range of sites with the revised models than with the original models. Identifying sites outside of the experience of a model is an important but often overlooked step in water-quality assessment. When the environmental conditions of an assessed site fall outside of the range of environmental conditions of the pool of reference sites used to train the model, the predictions will likely be inaccurate or imprecise. In this study, we used a nearest-neighbor approach to identify outliers, but other methods exist (e.g., Mahalanobis distances). The choice of method will ultimately depend on the goals of the assessment program and ease of implementation.

### Revised models alleviated shifting-baseline issues

Another goal was to minimize shifting-baseline issues, which may be of concern in water chemistry and other assessments (Gillon et al. 2016). Shifting baselines can occur when conditions at reference sites change over time because of human activity and those effects are not adjusted for (Pauly 1995). Nonstationary conditions can emerge when predictors associated with climate (e.g., previous year's precipitation), land use and cover (% urban development, % deciduous forest), and infrastructure (e.g., miles of roads in a watershed, number of dams) are used in models. Some water-chemistry models use nonstationary variables to predict how stream chemistry is expected to change under current climate conditions. For example, Olson and Cormier's national SC model (Olson and Cormier 2019) uses nonstationary predictors of climate (e.g., mean of monthly maximum temperature for the 2 mo before the sampling event) to predict temporal variation in SC levels. Although models that use temporally dynamic predictor variables may provide more precise predictions, which may be useful in some situations (e.g., to parse effects of land use from climate effects or to forecast future changes in water chemistry), routinely incorporating current climate conditions into models to predict reference-condition water chemistry can create potentially serious shifting-baseline issues. Although long-term climate metrics such as 30-y normals can alleviate shifting-baseline issues to some extent, they are not without drawbacks. Long-term averages will mask short-term variability in water chemistry that would better characterize the true range of naturally occurring variation expected under reference conditions, potentially limiting the responsiveness of assessments to dynamic environmental changes. This trade-off underscores the importance of carefully balancing temporal resolution and stability when incorporating climate predictors into models. Therefore, we suggest exercising caution with both approaches and considering the specific goals of the model application when setting benchmarks for reference-condition water chemistry.

### Using prediction limits to set water-quality benchmarks

In our study, the percentage of sites exceeding a benchmark varied from ~10 to ~30%, depending on the prediction limit used or whether site-specific or ecoregion-wide benchmarks were used. Upper prediction limits based on the SEE method were generally lower than those based on QRF, suggesting that benchmarks based on QRF could increase the risk of type II errors of inference, which contribute to continuing environmental degradation. However, the use of either method of site-specific modeling should lead to more appropriate benchmarks and lower uncertainty in their applicability than regionally set benchmarks (Hawkins et al. 2010, van Dam et al. 2019). For example, our results suggest that the SC benchmarks the USEPA uses for NRSAs for western ecoregions are not representative of the true distribution of reference-condition SC values at sites in the western United States. In our study, the 95th percentile of the distribution of SC values observed at reference sites was ~466 μS/cm, much lower than the benchmarks of 1000 to 2000 μS/cm that the NRSA program uses. Moreover, there is evidence that benchmarks ≥1000 μS/cm are likely too high to be protective of aquatic life, particularly for sensitive taxonomic groups. For example, mayfly abundance, drift, and metabolism can be strongly affected at SC levels ≤300 μS/cm (Clements and Kotalik 2016). The use of underprotective benchmarks could lead to considerable changes in community structure, the extirpation of species, and the further degradation of freshwater habitats (Clements and Kotalik 2016). Regional NRSA benchmarks for [TN] (range 249–1069 μg/L) and [TP] (range 41–127 μg/L) were more consistent with the 95th percentiles of the distribution of reference-site concentrations for [TN] and [TP], which were ~552 μg/L and 79 μg/L, respectively. However, our analyses imply that the ecoregion-derived benchmarks for [TN] and [TP] may be overprotective relative to either the SEE- or QRF-derived benchmarks. However, this result varied by ecoregion for both [TN] and [TP]. Our results suggest that ecoregion-derived benchmarks may be overprotective relative to site-specific benchmarks in some ecoregions (e.g., Western Mountains) but underprotective in others (e.g., Northern and Southern Plains).

### Management implications

We developed predictive models for SC, [TN], and [TP] based on a nationally available dataset that performed as well as, or better than, models based on GIS-derived predictor variables. Our revised models were anchored to a specific time period, which avoids shifting-baseline issues associated with nonstationarity in climate and should help management agencies detect water-quality impairment of stream and river reaches associated with both land use and climate change across the western United States. We suggest that similar site-specific models be used elsewhere to set water-quality benchmarks. However, the choice of what method to use when setting benchmarks should always be informed by the specific objectives of a project (e.g., whether benchmarks are to be used in a regulatory context or as a means of tracking water-quality trends), the ease of estimating benchmark values, and the ability to easily communicate methods to stakeholders, managers, and policymakers.

## ACKNOWLEDGEMENTS

Author contributions: JLC, DN, and CPH planned the project. DN performed the analysis and led the writing. JLC and CPH reviewed and commented on the manuscript.

We thank John Olson (California State University, Monterey Bay) for providing locations for new reference sites on which to train the revised models. We also thank 2 anonymous reviewers for their comments and suggestions, which greatly improved this manuscript. This work was supported by Cooperative Agreement L22AS00240 with the United States Bureau of Land Management. The SFS Endowed Publication Fund paid some of the publication costs of this paper (https://freshwater-science.org/publications/endowed-publication-fund).

Conflict of interest statement: The authors declare that they have no conflict of interest related to this manuscript. No financial or personal relationships that could influence the work presented have been identified.

## LITERATURE CITED

Akhtar, N., M. I. Syakir Ishak, S. A. Bhawani, and K. Umar. 2021. Various natural and anthropogenic factors responsible for water quality degradation: A review. Water 13:2660.

Baker, M. E., M. J. Wiley, P. W. Seelbach, and M. L. Carlson. 2003. A GIS model of subsurface water potential for aquatic resource inventory, assessment, and environmental management. Environmental Management 32:706–719.

BLM (Bureau of Land Management). 2015. AIM National Aquatic Monitoring Framework: Introducing the framework and indicators for lotic systems. Technical Reference 1735-1. National Operations Center, Bureau of Land Management, United States Department of the Interior, Denver, Colorado. (Available from https://www.blm.gov/documents/national-office/blm-library/technical-reference/aim-national-aquatic-monitoring-framework)

Breiman, L. 2001. Random forests. Machine Learning 45:5–32.

Clements, W. H., and C. Kotalik. 2016. Effects of major ions on natural benthic communities: An experimental assessment of the US Environmental Protection Agency aquatic life benchmark for conductivity. Freshwater Science 35:126–138.

Dewitz, J. 2023. National Land Cover Database (NLCD) 2021 Products: U.S. Geological Survey data release. (Available from https://doi.org/10.5066/P9JZ7AO3)

Dodds, W. K., W. H. Clements, K. Gido, R. H. Hilderbrand, and R. S. King. 2010. Thresholds, breakpoints, and nonlinearity in freshwaters as related to management. Journal of the North American Benthological Society 29:988–997.

Dodds, W. K., and R. M. Oakes. 2004. A technique for establishing reference nutrient concentrations across watersheds affected by humans. Limnology and Oceanography: Methods 2:333–341.

Fu, L., and Y.-G. Wang. 2011. Nonparametric rank regression for analyzing water quality concentration data with multiple detection limits. Environmental Science & Technology 45:1481–1489.

Genuer, R., J.-M. Poggi, and C. Tuleau-Malot. 2015. VSURF: An R package for variable selection using random forests. The R Journal 7:19–33.

Gibbons, R. D. 1987. Statistical prediction intervals for the evaluation of ground-water quality. Groundwater 25:455–465.

Gibbs, R. J. 1970. Mechanisms controlling world water chemistry. Science 170:1088–1090.

Gillon, S., E. G. Booth, and A. R. Rissman. 2016. Shifting drivers and static baselines in environmental governance: Challenges for improving and proving water quality outcomes. Regional Environmental Change 16:759–775.

Greathouse, E. A., J. E. Compton, J. Van Sickle. 2014. Linking landscape characteristics and high stream nitrogen in the Oregon Coast range: Red alder complicates use of nutrient criteria. Journal of the American Water Resources Association 50:1383–1400.

Hawkins, C. P., J. R. Olson, and R. A. Hill. 2010. The reference condition: Predicting benchmarks for ecological and water-quality assessments. Freshwater Science 29:312–343.

Hengl, T., M. Nussbaum, M. N. Wright, G. B. M. Heuvelink, and B. Gräler. 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 6:e5518.

Herlihy, A. T., and J. C. Sifneos. 2008. Developing nutrient criteria and classification schemes for wadeable streams in the conterminous US. Journal of the North American Benthological Society 27:932–948.

Hill, R. A., M. H. Weber, S. G. Leibowitz, A. R. Olsen, and D. J. Thornbrugh. 2016. The Stream-Catchment (StreamCat) dataset: A database of watershed metrics for the conterminous United States. Journal of the American Water Resources Association 52:120–128. (Available from https://www.epa.gov/national-aquatic-resource-surveys/streamcat-dataset#access-streamcat-data, Accessed March 2023)

Keiser, D. A., and J. S. Shapiro. 2019. Consequences of the Clean Water Act and the demand for water quality. The Quarterly Journal of Economics 134:349–396.

Le, T. D. H., M. Kattwinkel, K. Schützenmeister, J. R. Olson, C. P. Hawkins, and R. B. Schäfer. 2019. Predicting current and future background ion concentrations in German surface water under climate change. Philosophical Transactions of the Royal Society B: Biological Sciences 374:20180004.

Lek, S., M. Guiresse, and J.-L. Giraudel. 1999. Predicting stream nitrogen concentration from watershed features using neural networks. Water Research 33:3469–3478.

Liaw, A., and M. Wiener. 2002. Classification and regression by *randomForest*. R News 2:18–22. (Available from https://CRAN.R-project.org/doc/Rnews/)

Lin, J., J. E. Compton, R. A. Hill, A. T. Herlihy, R. D. Sabo, J. R. Brooks, M. Weber, B. Pickard, S. G. Paulsen., and J. L. Stoddard. 2021. Context is everything: Interacting inputs and landscape characteristics control stream nitrogen. Environmental Science & Technology 55:7890–7899.

Lintern, A., J. A. Webb, D. Ryu, S. Liu, U. Bende-Michl, D. Waters, P. Leahy, P. Wilson, and A. W. Western. 2018. Key factors influencing differences in stream water quality across space. Wiley Interdisciplinary Reviews Water 5:e1260.

Marcarelli, A. M., R. W. Fulweiler, and J. T. Scott. 2022. Nitrogen fixation: A poorly understood process along the freshwater-marine continuum. Limnology and Oceanography Letters 7: 1–10.

McKay, L., T. Bondelid, T. Dewald, J. Johnston, R. Moore, and A. Rea. 2012. NHDPlus version 2: User guide. Office of Water, United States Environmental Protection Agency, Washington, DC. (Available from https://www.epa.gov/waterdata/nhdplus-national-hydrography-dataset-plus)

Meinshausen, N. 2024. *quantregForest*: Quantile regression forests. R package version 1.3–7.1. (Available from https://CRAN.R-project.org/package = quantregForest)

Meybeck, M. 2004. The global change of continental aquatic systems: Dominant impacts of human activities. Water Science & Technology 49:73–83.

Ohlendorf, H. M., S. M. Covington, E. R. Byron, and C. A. Arenal. 2011. Conducting site-specific assessments of selenium bioaccumulation in aquatic systems. Integrated Environmental Assessment and Management 7:314–324.

Olson, J. R., and S. M. Cormier. 2019. Modeling spatial and temporal variation in natural background specific conductivity. Environmental Science & Technology 53:4316–4325.

Olson, J. R., and C. P. Hawkins. 2012. Predicting natural baseflow stream water chemistry in the western United States. Water Resources Research 48:W02504.

Olson, J. R., and C. P. Hawkins. 2013. Developing site-specific nutrient criteria from empirical models. Freshwater Science 32: 719–740.

Pauly, D. 1995. Anecdotes and the shifting baseline syndrome of fisheries. Trends in Ecology and Evolution 10:430.

Schwarzenbach, R. P., T. Egli, T. B. Hofstetter, U. von Gunten, and B. Wehrli. 2010. Global water pollution and human health. Annual Review of Environment and Resources 35:109–136.

Shen, L. Q., G. Amatulli, T. Sethi, P. Raymond, and S. Domisch. 2020. Estimating nitrogen and phosphorus concentrations in streams and rivers, within a machine learning framework. Scientific Data 7:161.

Smith, R. A., R. B. Alexander, and G. E. Schwarz. 2003. Natural background concentrations of nutrients in streams and rivers of the conterminous United States. Environmental Science & Technology 37:3039–3047.

Soranno, P. A., T. Wagner, S. L. Martin, C. McLean, L. N. Novitski, C. D. Provence, and A. R. Rober. 2011. Quantifying regional reference conditions for freshwater ecosystem management: A comparison of approaches and future research needs. Lake and Reservoir Management 27:138–148.

Stets, E. G., L. A. Sprague, G. P. Oelsner, H. M. Johnson, J. C. Murphy, K. Ryberg, A. V. Vecchia, R. E. Zuellig, J. A. Falcone, and M. L. Riskin. 2020. Landscape drivers of dynamic change in water quality of U.S. rivers. Environmental Science & Technology 54:4336–4343.

Stoddard, J. L., D. P. Larsen, C. P. Hawkins, R. K. Johnson, and R. H. Norris. 2006. Setting expectations for the ecological condition of streams: The concept of reference condition. Ecological Applications 16:1267–1276.

Suplee, M. W., A. Varghese, and J. Cleland. 2007. Developing nutrient criteria for streams: An evaluation of the frequency distribution method. Journal of the American Water Resources Association 43:453–472.

Toevs, G. R., J. J. Taylor, C. S. Spurrier, W. C. MacKinnon, and M. R. Bobo. 2011. Bureau of land management assessment, inventory, and monitoring strategy: For integrated renewable resources management. United States Department of the Interior, Bureau of Land Management, National Operations Center, Denver, Colorado. (Available from https://archive.org/details/assessmentinvent00toev/mode/2up)

USEPA (United States Environmental Protection Agency). 2000. Nutrient criteria technical guidance manual: Rivers and streams. EPA-822-B-00-002. Office of Water, Office of Science and Technology, United States Environmental Protection Agency, Washington, DC. (Available from https://www.epa.gov/nutrientpollution/nutrient-criteria-development-document-rivers-and-streams)

USEPA (United States Environmental Protection Agency). 2023. National rivers and streams assessment 2018–2019 technical support document. EPA-841-R-22-005. Office of Water: Office of Wetlands, Oceans and Watersheds, Office of Research and Development, United States Environmental Protection Agency, Washington, DC. (Available from https://www.epa.gov/national-aquatic-resource-surveys/national-rivers-and-streams-assessment-2018-19-technical-support)

van Dam, R. A., A. C. Hogan, and A. J. Harford. 2017. Development and implementation of a site-specific water quality limit for uranium in a high conservation value ecosystem. Integrated Environmental Assessment and Management 13:765–777.

van Dam, R. A., A. C. Hogan, A. J. Harford, and C. L. Humphrey. 2019. How specific is site-specific? A review and guidance for selecting and evaluating approaches for deriving local water quality benchmarks. Integrated Environmental Assessment and Management 15:683–702.

Vander Laan, J. J., and C. P. Hawkins. 2014. Enhancing the performance and interpretation of freshwater biological indices: An application in arid zone streams. Ecological Indicators 36: 470–482.

NWQMC (National Water Quality Monitoring Council), USGS (United States Geological Survey), and USEPA (Environmental Protection Agency). 2021. Water Quality Portal. (Available from https://doi.org/10.5066/P9QRKUVJ, accessed March 2023)

Weber, M. H., R. A. Hill, and A. F. Brookes. 2024. StreamCatTools: Tools to work with the StreamCat API within R and access the full suite of StreamCat and LakeCat metrics. (Available from https://usepa.github.io/StreamCatTools)

Webster, J. R., J. D. Newbold, and L. Lin. 2016. Nutrient spiraling and transport in streams: The importance of in-stream biological processes to nutrient dynamics in streams. Pages 181–239 *in* J. B. Jones and E. H. Stanley (editors). Stream ecosystems in a changing environment. Academic Press.

Yan, X., T. Zhang, W. Du, Q. Meng, X. Xu, and X. Zhao. 2024. A comprehensive review of machine learning for water quality prediction over the past five years. Journal of Marine Science and Engineering 12:159.

Zhou, M., Y. Zhang, J. Wang, Y. Shi, and V. Puig. 2022. Water quality indicator interval prediction in wastewater treatment process based on the improved BES-LSSVM algorithm. Sensors 22:422.

Zhu, M., J. Wang, X. Yang, Y. Zhang, L. Zhang, H. Ren, B. Wu, and L. Ye. 2022. A review of the application of machine learning in water quality evaluation. Eco-Environment & Health 1:107–116.