Freshwater Science, 2013, 32(3):719–740

DOI: 10.1899/12-113.1
Published online: 4 June 2013

Developing site-specific nutrient criteria from empirical models

John R. Olson[1] AND Charles P. Hawkins[2]

Department of Watershed Sciences, Western Center for Monitoring and Assessment of Freshwater Ecosystems, and the Ecology Center, Utah State University, Logan, Utah 84322-5210 USA

Abstract. Ecologically meaningful and scientifically defensible nutrient criteria are needed to protect the water quality of USA streams. Criteria based on our best understanding of naturally occurring nutrient concentrations would protect both water quality and aquatic biota. Previous approaches to predicting natural background nutrient concentrations have relied on some form of landscape categorization (e.g., nutrient ecoregions) to account for natural variability among water bodies. However, natural variation within these regions is so high that use of a single criterion can underprotect naturally occurring low-nutrient streams and overprotect naturally occurring high-nutrient steams. We developed Random Forest models to predict how baseflow concentrations of total P (TP) and total N (TN) vary among western USA streams in response to continuous spatial variation in nutrient sources, sinks, or other processes affecting nutrient concentrations. Both models were relatively accurate (root mean square errors <12% of the range of observations for independent validation sites) and made better predictions than previous models of natural nutrient concentrations. However, the models were not very precise (TP model: $r^2 = 0.46$, TN model: $r^2 = 0.23$). An analysis of the sources of variation showed that our models accounted for most of the spatial variation in nutrient concentrations, and much of the imprecision was caused by temporal or measurement variation. We applied 2 methods to determine upper prediction limits that incorporated model error and could be used as site-specific nutrient criteria. These site-specific candidate nutrient criteria better accounted for natural variation among sites than did criteria based on regional average conditions, would increase protection for streams with naturally low nutrient concentrations, and specified more attainable conditions for those streams with naturally higher nutrient concentrations.

Key words: models, nitrogen, nutrient concentrations, nutrient criteria, phosphorus, Random Forests, reference condition, streams, United States, water quality.

Nutrients are among the most important stressors to aquatic ecosystems and lead to eutrophication of local water bodies and downstream lakes and estuaries. Nutrient pollution has increased markedly over the last 50 y, with >50% of stream and 78% of coastal waters now exhibiting eutrophication (USEPA 2011). To prevent further harm and to set standards for restoration, the US Clean Water Act requires that criteria be established to protect the designated uses of each water body. Criteria can be in narrative or numeric form, but the USEPA has long recommended use of numeric nutrient criteria to identify the level of impairment, prioritize water bodies for management, and set remediation goals for individual water bodies (USEPA 2011). Numeric criteria based on naturally occurring nutrient concentrations would protect all uses, but are especially important for protecting those aquatic biota that are adapted to specific trophic conditions associated with different nutrient environments (Dodds 2007). Criteria also should be applicable at the scale used for water-quality management, the individual water body. The challenge in establishing such criteria lies in estimating the naturally occurring, reference concentrations expected at individual water bodies.

Several approaches have been developed to predict background nutrient conditions and define criteria. One approach is to base a criterion on some percentile value of the distribution of nutrient concentrations observed at reference sites within a region (e.g., 75th percentile, USEPA 2000; 86th percentile, Suplee et al. 2007). Another is to model background nutrient concentrations as a function of ecoregion, runoff, and atmospheric deposition (for N) or in-stream loss (for P) (Smith et al. 2003). In a 3rd approach, Dodds and Oakes (2004) modeled nutrient concentrations as a function of landuse disturbance within separate

[1] E-mail addresses: john.olson@usu.edu
[2] chuck.hawkins@usu.edu

ecoregions and depended on the ecoregions to control for natural variation. They predicted naturally occurring concentrations by applying the model with disturbance set to 0 at altered sites because disturbance was used as a predictor in the model. These approaches all control for natural variation in nutrient concentrations caused by differences in geology, climate, or vegetation by classifying sites into nutrient ecoregions that separate sites spatially into groups based on environmental similarities. However, whether regionalizations control sufficiently for natural variation in water chemistry and other ecosystem attributes is questionable (Hawkins et al. 2010).

Even when landscape classifications are based on known environmental drivers, they often account for insufficient amounts of natural variation in nutrient conditions to allow prediction of expected natural nutrient concentrations. Herlihy and Sifneos (2008) concluded that the 14 nutrient ecoregions covering the contiguous USA do not control natural variability well enough to allow establishment of regional criteria, specifically in the Pacific Northwest. Total P (TP) and total N (TN) concentrations varied $\geq 3\times$ even among reference sites within some of the finer-resolution level-III ecoregions (85 regions for the contiguous USA; fig. 5 by Herlihy and Sifneos 2008). Cheruvelil et al. (2008) found that multiple regionalization schemes were ineffective in partitioning natural variation in TP and TN among minimally disturbed lakes in Michigan. Robertson et al. (2006) noted several inherent problems in using ecoregions to account for variation, including the difficulty of developing a single classification that adequately parses natural variation of multiple chemical constituents when each constituent responds to a different set of processes. They also noted that ecoregions are often confounded with land use because human development occurs disproportionately in ecoregions with favorable environmental attributes. For example, if the amount of agriculture is correlated with natural differences in soil and vegetation type, then regions delineated based on soils or vegetation are likely to differ in water chemistry because of differences in land use and variation in natural features. Identifying appropriate background concentrations in streams that flow across multiple regions and assigning criteria to such streams is problematic (Dodds and Oakes 2004).

Other investigators have tried using typological or reach-level classification approaches to control for natural variation in nutrient concentrations (Snelder et al. 2004, Robertson et al. 2006, Herlihy and Sifneos 2008). These typologies were more effective than ecoregions, but nutrient concentrations still varied up to an order of magnitude within some classes. Many of the environmental drivers of water chemistry vary continuously (e.g., climate, topography, vegetation), so any discrete classification imposed on these gradients must contain a certain amount of within-class variation.

If large amounts of unexplained natural variation occur within landscape or waterbody classes, establishing criteria that are both attainable and protective across the range of expected conditions is difficult. Any criterion chosen from across a large range of possible natural conditions will be underprotective for some sites and overprotective for others. An example of underprotection would be a site with very low natural nutrient concentrations in a highly variable region with a criterion significantly higher than that site's natural background condition. Such a site would have to be altered substantially before the nutrient concentrations violated the criterion and prompted action. Ice and Binkley (2003) described an example of overprotection in which the nutrient concentrations found in 3 streams draining undisturbed forest watersheds would exceed regional criteria, a result indicating that these criteria were set too low. They concluded that "Water quality standards will be acceptable only when they reflect what is physically achievable…" (Ice and Binkley 2003, p. 27). Given the monetary and societal costs associated with restoring nutrient-enriched streams, it is critical that management decisions be guided by achievable and reliable criteria.

Nutrient criteria should be based on the best estimates of expected natural or near-natural conditions, but making these estimates is difficult given the complex environmental processes that influence nutrient concentrations. Smith et al. (2003) developed regression models to predict natural background nutrient concentrations, but because they lacked access to information on vegetation, soils, or geology, they relied on ecoregions to account for all of these environmental effects. Ice and Binkley (2003) noted that although ecoregions explain some variation in nutrient concentrations, they do not account for the influence of finer-scale factors, such as geology or forest type. Dodds and Oakes (2004) called for consideration of spatially variable characteristics, such as geology, slope, and drainage area, to better account for natural variation in water chemistry within ecoregions. New spatial data describing environmental factors that can influence water chemistry have been produced (Olson and Hawkins 2012), and new modeling techniques that account for nonlinear and interacting predictors have been developed (e.g., Random Forests and Artificial Neural

Networks). These advancements in data and modeling provide an opportunity to develop models in which stream nutrient concentrations are predicted as joint functions of potential nutrient sources and sinks without the need to rely on spatial classifications like ecoregions.

Our main objectives were to develop models to predict baseflow nutrient concentrations for individual stream reaches and then to identify site-specific nutrient criteria based on these model predictions. We first describe how we modeled site-specific variation in naturally occurring TN and TP concentrations. Even the most pristine streams receive some atmospheric deposition from both natural and anthropogenic sources, so we did not attempt to parse the effects of anthropogenic deposition from estimates of naturally occurring nutrients. We then describe 2 methods for estimating prediction error and demonstrate how these methods can be applied to estimate the highest probable naturally occurring nutrient concentration at a site, i.e., a candidate site-specific nutrient criterion.

Methods

Nutrient concentration data

We assembled a data set of TP and TN concentrations from samples collected during baseflow conditions by multiple agencies from 823 reference-condition streams across the western USA (Fig. 1, Table 1). Samples were collected from wadeable streams from almost all environments occurring in the western USA including mountains, deserts, coastal areas, and the Great Plains. Baseflow conditions at the time of sampling were either determined from gauge records or were verified in the field by individual crews. Sample TP and TN concentrations were measured from unfiltered grab samples by persulfate oxidation and colorimetry (TP and TN) or calculated as the sum of total Kjeldahl N plus NO_3^- and NO_2^- (TN). Laboratory precision was not available for all data, but detection limits were generally 2 to 10 μg/L for TP and 10 to 60 μg/L for TN. We used concentrations derived from individual grab samples instead of long-term averages or estimates of nutrient loads despite the noisiness of this type of data (Knowlton and Jones 2006) because most regulatory agencies use estimates from grab samples in their assessment programs. Also, the number of sites with grab-sample data far exceeded the number of sites that had the frequent, multiple measurements needed to calculate loads. The data from many grab samples allowed us to develop models whose scope included a broad range of environments. Sites were originally identified as being in reference condition by the sampling agency, but to ensure consistency, we screened sites to verify that their catchments had little to no human disturbance except for atmospheric deposition (i.e., all sites had <10% agriculture or urban land use, and 95% of sites had <2% of either land use; see Olson and Hawkins 2012).

Environmental predictors

We used a geographic information system (GIS) to measure spatial variation in factors potentially affecting nutrient concentrations among sites. These factors include direct effects associated with spatial variation in sources (e.g., rock P, N deposition) and sinks (e.g., P deposition in lakes, removal of N by denitrification). We also measured factors that could indirectly affect nutrient concentrations (e.g., factors associated with evaporation or aquatic and terrestrial nutrient processing rates) and temporal data describing seasonal changes in climate or vegetation. Our measurements of spatial data included average upstream catchment conditions and the value of each variable at the sampling point. We delineated catchments by applying the Multi-Watershed Delineation Tool (Chinnayakanahalli 2006) to 30-m Digital Elevation Models. In total, these measurements produced 182 potential predictor variables for each site. We describe the major categories of predictors and the specific predictors selected for the final models below. The full list and descriptions of predictors is available in Table S1 (available online from: http://dx.doi.org/10.1899/12-113.1.s1).

Data on potential sources of P and N include descriptions of underlying geology, amounts of atmospheric deposition, and distributions of N-fixing plants. All geologic assessments were derived from the Preliminary Integrated Geologic Map Databases for the USA (Ludington et al. 2007, Stoeser et al. 2007). Basalts can be sources of elevated stream P (Meybeck 1982), so we measured the % of each catchment underlain by volcanic rocks. We also measured the average bedrock composition of P_2O_5, N, CaO, MgO, and S in each catchment (see Olson and Hawkins 2012 for details). We estimated the amount of bedrock NH_4^+ because bedrock N in the form of NH_4^+ is more easily weathered than organic forms (primarily kerogen; Holloway and Dahlgren 2002). NH_4^+ exists in other rock types, but we based our estimates of NH_4^+ rock content only on metamorphic rocks because mineralization of N is associated with diagenesis and metamorphism (Holloway and Dahlgren 2002). We extracted bedrock N values from all geologic map units associated with metamorphic

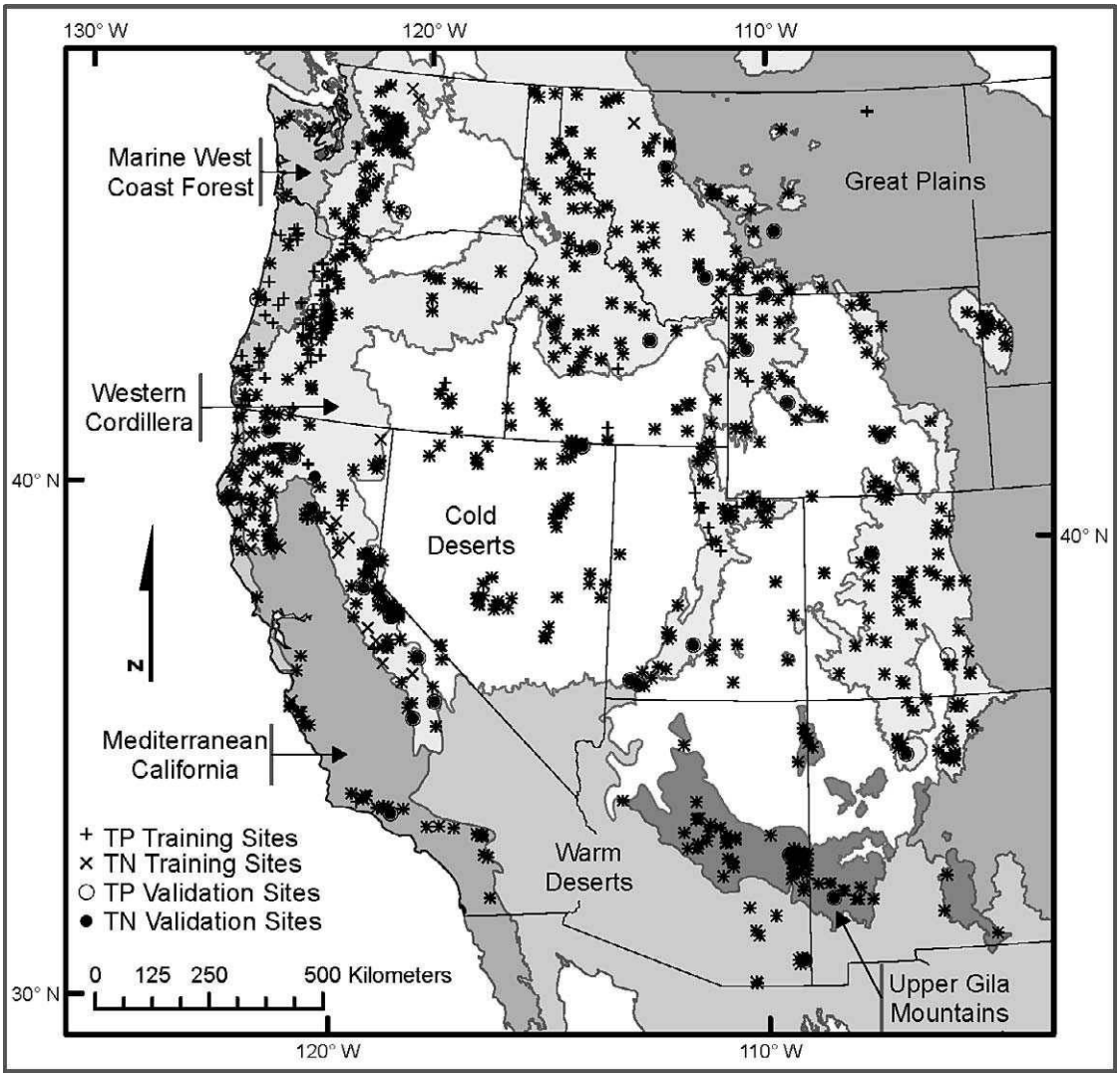


FIG. 1. Map of 782 training and 41 validation sites sampled for total P (TP), total N (TN), or both by ecoregion and state.

rocks and applied this value as our estimate of bedrock NH_4^+ concentration. Atmospheric deposition was measured as the long-term (1994–2006) average wet-deposition concentrations of NO_3^-, Ca, Na, and SO_4^{2-} from the National Atmospheric Deposition Program National Trends Network. Dry deposition can be a major source of N, so we also estimated catchment average annual dry + wet TN deposition.

TABLE 1. Sources of water chemistry data. US EPA = US Environmental Protection Agency, USGS = US Geological Survey.

Data source	No. sites	Years collected	Location/contact
Arizona Department of Environmental Quality	25	1994–2008	Patrice Spindler
California Department of Fish and Game	46	2003–2008	Andrew Rehn
Eastern Sierra Nevada Dataset	22	2000–2002	Dave Herbst
US EPA Environmental Monitoring and Assessment Program	337	2000–2004	http://www.epa.gov/emap2/
USGS National Water-Quality Assessment Program	41	1973–2008	http://water.usgs.gov/nawqa/
New Mexico Environment Department	25	1999–2007	Shann Stringer
Oregon Department of Environmental Quality	67	1992–2002	Shannon Hubler
Utah State University	255	2001–2003	John Olson
USGS National Water Information System	5	1981–1995	http://waterdata.usgs.gov/nwis

We obtained these estimates by applying the Watershed Deposition Tool to calculate long-term average deposition from output of the Community Multiscale Air Quality model (CMAQ; Schwede et al. 2009). N-fixing plants can be the dominant source of N in some streams (e.g., Compton et al. 2003), so we developed several predictors describing the potential distribution of N-fixing woody plants identified by the US Department of Agriculture PLANTS Database as naturally occurring in the western USA. These plants included *Alnus incana*, *Alnus rubra*, *Ceanothus velutinus*, and *Prosopis glandulosa*. We used the LANDFIRE Biophysical Settings Model descriptions and layers, which together describe presettlement vegetation patterns, to develop maps of the potential distributions of these species under natural conditions. First, we identified which LANDFIRE Biophysical Settings Model descriptions listed each species as either occurring or dominant (LANDFIRE 2011b). Then, we extracted those grid cells associated with the identified Biophysical Settings Model from the LANDFIRE Biophysical Settings layer (LANDFIRE 2011a) to create layers describing the expected locations at which each species would be either present or dominant in our study area. We also calculated % cover of *A. rubra* for each catchment from estimates of current forest composition derived from Gradient Nearest Neighbor imputation (Ohmann et al. 2007) of areas across the Pacific Northwest by the Landscape, Ecology, Modeling, Mapping, and Analysis project (LEMMA 2011).

Potential sinks for nutrients include uptake or retention by vegetation, soils, lakes, or wetlands; denitrification; and chemical precipitation or adsorption. To characterize spatial differences in potential vegetative uptake, we used long-term (2000–2009) average MODIS satellite Enhanced Vegetation Index (EVI) values (Huete et al. 2002) as a proxy for spatial variation in plant biomass. Because MODIS EVI data are available in weekly increments starting in 2000 we could potentially have used it to characterize temporally specific differences in vegetative uptake (i.e., EVI for the specific time of the sample or increase in EVI in the previous month). However 10% of our data were collected before MODIS became operational, so we relied on day of year of the sample to account for seasonal variations in vegetative uptake. We characterized major differences in vegetation composition with data from the 2001 National Land Cover Dataset (NLCD; Homer et al. 2004). We used maps of soil organic C (SOC; Global Soil Data Task Group 2000) and soil organic matter (SOM; NRCS 2011) content to characterize the potential release or immobilization of nutrients by soils caused by microbial uptake or chelation associated with SOC or SOM. We described potential differences in nutrient retention by lakes and wetlands by measuring the % of each catchment classified as lake, wetland, or both (i.e., water body) in both the NLCD and the National Hydrography Dataset (NHD; USGS 2006). We assessed the size of the largest water bodies in each catchment and the amount of flow routed through these water bodies in the NHD data. We also measured environmental variables associated with differences in conditions favorable to denitrification, such as soil bulk density (lower pore connectivity with increased density creates more anaerobic sites) or the amount of surface–subsurface hydrologic exchange in streams (increased exchange brings more N in contact with hyporheic waters). We obtained soil bulk density from the US General Soil Map (NRCS 2011). We characterized surface–subsurface hydrologic exchange by both average catchment hydraulic conductivity and an index of groundwater velocity estimated with the MRI-Darcy model (Baker et al. 2003). The MRI-Darcy model applies Darcy's equation within a GIS environment (see Olson and Hawkins 2012 for details). We also measured other factors that could potentially influence chemical precipitation or adsorption of nutrients where spatial data were available. These variables included the amount of Ca available from either bedrock or atmospheric sources that could act as a coprecipitate with P, and soil pH, which could influence adsorption or cation exchange.

We used long-term estimates (1971–2000) of average precipitation, number of wet days, air temperature, day of last freeze, and relative humidity produced by the Parameter-elevation Regression on Independent Slopes Model (PRISM; Daly et al. 1994) to estimate the effects of dilution and evaporative concentration. Temporal variation in precipitation can influence nutrient concentrations, so we also measured PRISM monthly mean precipitation for the month of the sample, mean precipitation for the month previous to the sample, and mean annual precipitation for the year previous to the sample.

We also measured other factors that could potentially affect processing rates or retention or that could act as proxies for factors we could not measure. These variables included soil order and properties (e.g., available water content, erosion factor, and % hydric soils), topography (e.g., elevation, relief, and catchment shape), catchment area, Level II ecoregion, and average channel slope. We also included measurements of other atmospheric deposition components

not directly related to nutrient concentrations, e.g., Mg, Na, Cl, and SO_4^{2-}.

Model development and evaluation

We used the nonparametric modeling technique Random Forest (RF; Breiman 2001) to develop empirical predictive models. RF models outperform multiple linear regression models for other water-chemistry constituents because of their ability to account for interactions between variables and nonlinear relationships (Olson and Hawkins 2012). RF models are ensembles of classification and regression trees (CART; Breiman et al. 1984). Observations are recursively split into groups, minimizing the remaining unexplained variance within each group. Splits are constructed as a series of binary rules based on one of the explanatory variables. CART models are sensitive to small changes in training data, but RF overcomes this limitation by growing multiple individual trees using a bootstrap sample of the training data and a random sample of the predictors at each split. RF predictions are then generated by averaging the predictions of all trees. RF estimates the predictive accuracy of the model from observations that were excluded from each bootstrap sample (out-of-bag error) and the importance of each predictor by measuring how out-of-bag error changes when each predictor is permuted. We implemented RF with the R package *randomForest* (Liaw and Wiener 2009) to create 1500 trees for each model. Prediction errors in individual trees caused by overfitting cancel each other when averaged over large numbers of trees constructed from random subsets of both data and predictors, so the resulting RF prediction does not overfit the data even when a large number of predictors is used (Breiman 2001). However, use of parsimonious models and limiting the number of predictor variables to be calculated are still desirable. To create the most parsimonious model and to minimize the number of correlated predictors, we modeled iteratively and removed correlated or low-importance predictors until a model's out-of-bag error began to increase. We used partial-dependence plots to visualize relationships between nutrient concentrations and predictors, and we removed any predictors for which the direction of response in nutrient concentrations changed $>3\times$ because such patterns are likely to be spurious relationships. After predictor variables were selected, we used the *tuneRF* function to optimize the size of the random sample of the predictors tried at each split. We corrected for a small bias inherent in RF regression models (Zhang and Lu 2012) by applying the bias-correction function internal to the *randomForest* package.

We used the training (internal) data and an external validation data set to evaluate model performance. We selected external validation data prior to model development by randomly sampling 5% of sites, stratified by level II ecoregion (CEC 2006) to ensure that the validation set represented all environments. Internal evaluations were based on out-of-bag observations (analogous to cross validation), which allowed us to assess how well the models performed across the widest range of conditions. External validation allowed us to assess rigorously the applicability of these models to completely independent observations. We quantified model performance with the Nash–Sutcliffe Model Efficiency coefficient (NSE) and r^2 values associated with linear regressions of observed vs predicted concentrations (Piñeiro et al. 2008). We used an equivalence test to assess model bias (systematic over- or underprediction) and consistency (deviance between observations and predictions remains constant over their ranges) by testing whether the regression of observed vs predicted concentrations had an intercept = 0 and a slope = 1 (Robinson et al. 2005). Intercepts $\neq 0$ indicate model bias, whereas slopes $\neq 1$ indicate that model predictions lack consistency across the range and the model over- or underpredicts at the extremes. The equivalence test approach reverses the test from a null hypothesis of agreement between observations and predictions to a null hypothesis of less than a given difference. This test shifts the burden of proof to the model, and rejection of the null hypothesis indicates predictions are sufficiently similar to the observations for that particular application. A failure to reject the null hypothesis (assessed with $\alpha = 0.05$) indicates either insufficient evidence of a similarity between predictions and observations or a true difference. We considered slopes ranging from 0.75 to 1.25 and intercepts ranging from -0.25 to 0.25 (i.e., region of equivalence) to be sufficiently similar based on previous applications of this method by others. Instead of applying the equivalence test once, we used a bootstrap analysis with 10,000 resamples of predictions and observations to estimate the proportion of results that would fall within the region of equivalence for both intercept and slope. We also used the Root Mean Square Error (RMSE) to assess model accuracy. Last, we compared the performance of our model with the only other models that predict background nutrient concentrations across the western USA, the empirical models developed by Smith et al. (2003). Smith et al. used regression to predict transport of nutrients into streams and used the

SPARROW model to predict nutrient losses during stream transport. These combined models predicted flow-weighted nutrient concentrations for individual reaches, and Smith et al. suggested that these predictions or their regional frequency distributions could assist in development of nutrient criteria. We compared the Smith et al. model with ours by extracting predictions of flow-weighted concentrations for each stream reach (shown in fig. 7 by Smith et al. 2003; available at http://water.usgs.gov/nawqa/sparrow/intro/Smithetal_ES&T_2003_fig7.xls) for which we had a corresponding validation sample.

Our predictors primarily describe static spatial variation among sites, but we also wanted to assess how much variation in nutrient concentrations potentially could be attributed to temporal or measurement variation. We assessed the magnitude of temporal or measurement variation in concentrations by calculating the ratio of between-site variance (spatial signal) to within-site variance related to temporal and measurement noise, i.e. the signal-to-noise (S:N) ratio (Kaufmann et al. 1999). For example, if more variation existed among multiple sites than existed among all repeated samples from the same sites, then the S:N ratio would be high. We used these S:N ratios to estimate the best possible r^2 that static predictors could produce. We estimated variance among within-site replicate samples from a subset of 41 US Environmental Protection Agency (EPA) Environmental Monitoring and Assessment Program (EMAP) and Utah State University (USU) sites sampled multiple times for both TP and TN. These samples exhibited temporal variation comparable to that seen by Chételat and Pick (2001). We added all of these replicate samples to our original training data set and derived the variance within and among sites from a linear mixed model built with the R package *lme4*. The model treated sites as a random effect and did not contain any fixed effect. We calculated the S:N ratio from these 2 variances and the maximum possible r^2 value as: $\max(r^2) = \text{S:N}/(\text{S:N} + 1)$ (J. Van Sickle, USEPA Western Ecology Division, personal communication, illustrated in fig. 2 by Stoddard et al. 2008). We calculated among-site variance with data from all sites instead of only those sites with replicate samples because this larger data set provides a more representative estimate of the natural variation in stream nutrient concentrations across the western USA.

Determining highest probable concentrations based on model predictions

Site-specific nutrient criteria should incorporate both the model prediction of nutrient concentrations and prediction uncertainty arising from unaccounted variation, imperfect model structure, and error in measuring predictor values and nutrient concentrations. Prediction uncertainty can be quantified by establishing a prediction interval describing the range of conditions expected at a site. The upper prediction limit (PL) of this interval establishes the upper limit of the expected nutrient condition and accounts for prediction uncertainty arising from unexplained variation and model uncertainty. Distribution-based statistical methods (e.g., linear regression) can produce prediction intervals from an assumed normal distribution, but nondistributional methods like RF cannot. Quantile Regression Forests have been proposed as a method for determining prediction intervals (Meinshausen 2006), but this approach has 2 shortcomings. RF models cannot extrapolate beyond the range of the data used to construct them, so quantiles based on RF models become constrained at the lower and upper ends of the data. Also, the quantiles produced by quantile random forest models do not account for uncertainties associated with the estimates of a given quantile. We relied instead on 2 forms of empirically derived prediction intervals to develop reliable prediction intervals for our RF models.

The Simple Empirical Error (SEE) method empirically determines the amount of error for each prediction from a bootstrap sample of residuals from the training data (J. Van Sickle, personal communication). For each prediction, we sampled all residuals 500 times with replacement and added each sampled residual to the prediction to create an empirical distribution of the prediction plus error. We selected the 95th percentile of this distribution as the upper PL for that prediction.

The 2nd method is a variation of the UNcertainty Estimation based on Local Errors and Clustering (UNEEC) method of Shrestha and Solomatine (2008). UNEEC is similar to SEE in that errors are determined from a bootstrap sample of residuals from the training data, but instead of using a sample of all residuals, UNEEC uses residuals from only those samples similar to the site we are trying to predict. Sample residuals for similar sites were derived by first clustering all training observations by their environmental properties and then bootstrap-sampling the residuals of each cluster and selecting the 95th percentile as the error for that cluster. For each prediction, probability of membership in each cluster is used to calculate a weighted average of the 95th percentile errors for all clusters. This weighted-average error is then added to the prediction to determine the upper 95th percentile PL for that

prediction. We created clusters based on those environmental variables selected for the RF model. These environmental data were first standardized to a common scale and then clustered (k-means clustering). We selected the number of clusters to minimize the sum of squares and to ensure the minimum number of samples included in each cluster was >50. We then randomly sampled the residuals of the training data for each cluster 500 times with replacement and determined the 95th percentile value. We assessed probability of cluster membership for new observations by applying a separate RF model built with the same transformed environmental variables used in clustering. We used the probabilities of cluster membership as weights when calculating the average 95th percentile error to be added to each prediction to determine the upper PL.

Results

Model structure and performance

Relationships between nutrient concentrations and most predictors were consistent with our understanding of how the natural environment influences nutrient concentrations (Figs 2, 3). Both models included factors related to sources and sinks, but 2 predictors, both related to geologic sources, in the TP model were clearly more important than the others. The TN model did not include any clearly dominant predictors, and TN was almost equally influenced by predictors related to sources and sinks. The TP predictors were almost entirely static (with the exception of previous year's precipitation), whereas the TN model included temporal measures like day of year and precipitation during the 2 mo prior to sampling.

We tried to eliminate correlated variables during variable selection, but in several cases, removing correlated predictors degraded model performance. To maximize the model's ability to make predictions, we retained correlated variables if they improved model performance. The only predictors in our TN model that were strongly correlated were atmospheric SO_4^{2-} and NO_3^- deposition ($r = 0.9$). Correlated TP predictors included relative humidity and SOC ($r = 0.8$), relative humidity and atmospheric Ca deposition ($r = 0.64$), relative humidity and previous year's precipitation ($r = 0.63$), SOC and previous year's precipitation ($r = 0.67$), local minimum temperature and EVI ($r = 0.63$), % volcanic lithology and rock P concentration ($r = 0.69$), and soil erosion factor and soil water capacity ($r = 0.61$). RF models are robust to the effects of correlated predictors (Cutler et al. 2007). However, correlated predictors can cause variable importance measures to be unreliable (Strobl et al. 2008), so inferences regarding the relative importance of different processes in Figs 2 and 3 should be made with caution.

Both models predicted nutrient concentrations without significant bias, but were relatively imprecise (Table 2, Fig. 4A, B). The TP model accounted for <½ of the variation in TP concentrations, and the TN model accounted for <⅓ of the variation in TN concentrations. However, both models did have positive, if modest, NSEs indicating some predictive power. RMSEs of both models were <12% of the range of observed values (TP range: 1–192 μg/L, TN range: 5–960 μg/L). Model fit varied slightly between training and validation data, but we saw no evidence that the RF models were overfit to the training data. Only the TP model showed any evidence of bias, which was only slight (−2.3 μg/L) with 16% of the bootstrapped validation samples having an intercept less than the specified region of equivalence. Both models had slopes equivalent to 1 when assessed with training data but not when assessed with validation data, results indicating that predictions were not always consistent with observed values at new locations. For validation data, 51% of the bootstrap slope estimates for the TP model fell above the region of equivalence (Fig. 4A). The slope of all predictions together was 1.3, but this result was heavily influenced by the single validation observation >100 μg/L. This slope > 1 indicates that the model increasingly overpredicted with increasing TP concentrations. The equivalence test for slope showed the opposite pattern for the TN model, with 64% of the bootstrap estimates of slope falling below the region of equivalence and a smaller slope (0.66), indicating underpredictions at higher concentrations (Fig. 4B). Both models explained much more variance in nutrient concentrations than did predictions based on the Smith et al. (2003) models (Table 2), which, given their negative NSEs, predicted nutrient concentrations no better than the mean of the data.

Our models had relatively low r^2 values, but the results of our S:N analysis indicated that both models explained most of the spatial variation in nutrient concentrations (Table 3). The remaining unexplained variation was the result of either temporal variation or measurement error. To assess how much of the explainable spatial variation was accounted for by spatial predictors, we removed temporal predictors from both models by removing day of year from the TN model and replacing temporal precipitation measures with long-term average precipitation in both models. The spatial-only TP model had an r^2 of 0.39 and accounted for 59% of the static spatial

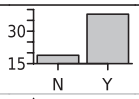
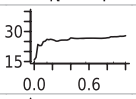
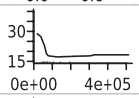
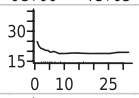
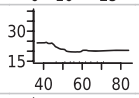
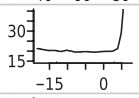
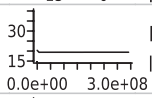
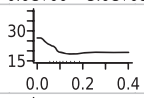
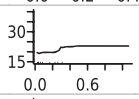
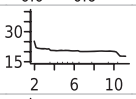
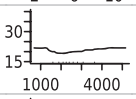
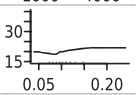
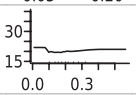
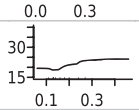
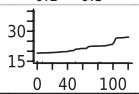
Predictor	Imp	Effect	Interpretation
Gila Mtns. ecoregion	36 (0.8)		Unusual combination of steep topography and large amounts of young basalts in Upper Gila Mountains/Mogollon Rim result in increased P (Rampe et al. 1981)
% volcanic lithology	31 (1.8)		Volcanic rocks are large source of P (Dillon and Kirchner 1975), that chemically weather more quickly than other igneous rocks types (Gislason et al. 1996)
Previous year's precipitation	26 (1.6)		Captures both spatial variability in amount of dilution with increasing discharge and the wash out of retained P by previous year's flood events (House 2003)
Rock % CaO	24 (1.3)		Increased Ca availability results in co–precipitation with P in river bed (House 2003) although contrasts with increased rock weathering (Mulholland 1992, Cross and Schlesinger 2001)
Relative humidity	24 (1.5)		Low humidity results in increased evaporative concentration of solutes, in addition to affecting TP by decreasing soil organic carbon
Local minimum temperature	22 (0.9)		Lack of freezing results in increased water flow and less nutrient uptake (Green and Finlay 2010), more shading and less uptake by periphyton, or less P sorption by wetlands (Wang et al. 2007)
Area largest water body	21 (1.0)		Presence of lakes or wetlands acts as a sink for P in catchment, with larger water bodies retaining more P (Smith et al. 2003)
Average channel slope	21 (0.9)		Low slopes have greater hydrologic flushing of P from saturated surface soils (D'Arcy and Carignan 1997). Lower P retention on high slopes (Hill et al. 2010) may result in less P export at baseflow
Atmospheric Ca deposition	21 (1.5)		Ca deposition acts as surrogate for dust deposition, the major source for atmospheric P (Reynolds et al. 2001)
SOC	20 (2.8)		High SOC increases biotic P uptake (Kirschbaum 2000) and immobilization in organic form (Stevenson and Cole 1999), but also correlated with old leached soils (Walker and Syers 1976)
EVI	19 (1.1)		Initially increasing vegetation retains P resulting in less P in streams, but at higher levels additional vegetation results increases chemical weathering and release of P
Soil water capacity	19 (1.0)		Higher water capacities associated with fine soils (i.e., not sandy soils) are correlated with higher surface runoff and erosion rates (Panagopoulos et al. 2007)
Soil erosion K factor	18 (1.3)		Highly erodible soils transport greater amounts of P with suspended sediment
Rock % P	16 (1.1)		High rock P acts as source of P within catchment, % rock P captures variation in P among non–volcanic rocks
% alfisols	15 (1.1)		Alfisols may increase P export by providing Fe allowing P to bind to dissolved humic matter (Dillon and Molot 1997) or reduce P retention in upper horizons due to low clay content (Bhadha and Jawitz 2010)

FIG. 2. Predictors, relative importance (Imp), direction of effect, and associated mechanisms for total P (TP) model. Importance is the % increase in mean squared error when the predictor is removed with standard error of the mean in parentheses (calculated from 50 separate models). Effect is illustrated as partial dependence plots of each predictor with all other predictors held constant. Change in predictor is displayed on the *x*-axis and change in TP is displayed on the *y*-axis. Mtns = mountains, N = no, Y = yes, SOC = soil organic C, EVI = enhanced vegetation index.

variation in concentrations, i.e., the model explained 39% of the observed variation compared with a maximum possible of 66%. The spatial-only TN model had an r^2 of 0.28 and accounted for 52% of the spatial variation.

Determining the highest probable concentration based on model predictions

The SEE and UNEEC methods produced similar upper PLs (Fig. 5A, B). Each method produced site-specific

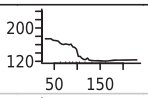
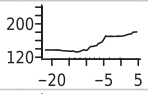
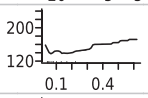
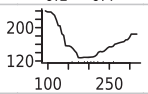
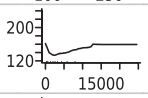
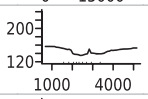
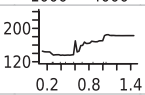
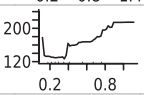
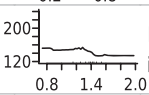
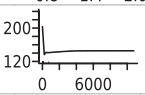
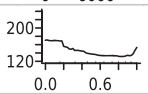
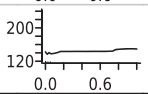
Predictor	Imp	Effect	Interpretation
Mean number of wet days	27 (0.9)		Wet soils resulting from more wet days increases denitrification (Bollmann and Conrad 1998), more precipitation increases dilution, dry conditions favor N build up and flushing (Van Miegroet et al. 2001)
Minimum temperature	25 (0.7)		Higher temperatures increase N–fixation in soil (Cleveland et al. 1999) and streams (Marcarelli and Wurtsbaugh 2006) and also litter decomposition (Park et al. 2003) and soil nitrification
Atmospheric Na deposition	24 (0.9)		Mechanism appears to be increased release of NH_4 associated with Na (and then nitrification of the released NH_4), not direct exchange with Cl (J ana Compton, USEPA–Corvallis, Oregon, personal comm.)
Day of year	24 (0.8)		Phenology related uptake of N by plants, reaching its maximum in summer (Also note higher inputs in spring than fall from snowmelt)
Prior 2 mo precipitation	23 (0.9)		Precipitation favors N fixation, litter decomposition (Lewis et al. 1999), and flushing (Kane et al. 2008) in mesic areas, whereas in xeric areas it may increase plant uptake and/or denitrification
Atmospheric NO_3 deposition	21 (0.8)		Source of N from anthropogenic and natural sources
Atmospheric SO_4 deposition	21 (0.7)		No known mechanism, but Likens et al. (2002) and Cai et al. (2011) saw same pattern, may be surrogate for other source (like dry N deposition), or perhaps SO_4 displaces NO_3 adsorbed to soils
EVI	20 (1.1)		Increasing biomass results in N uptake to a point, where forest maturation results in decreased N retention
Soil bulk density	18 (1.0)		Increasing soil density creates more anaerobic sites, therefore increases denitrification in soils (Torbert and Wood 1992)
Ground water index	16 (0.9)		Higher ground water index could reflect either up welling of ground water with little soil contact or increasing hyporheic zone contact increasing denitrification (Grimaldi and Chaplot 2000)
% evergreen	15 (0.8)		Kane et al. (2008) saw same pattern, perhaps due to slower processing or lower N content of evergreen litter (Lopez et al. 2001, Washburn and Arthur 2003), or as surrogate for a soil attribute
% *Alnus rubra* dominated	10 (0.8)		*Alnus rubra* (where present) is a major source of N (Compton et al. 2003)

FIG. 3. Predictors, relative importance (Imp), direction of effect, and associated mechanisms for total N (TN) model. Importance is listed as the % increase in mean squared error when the predictor is removed with standard error of the mean in parentheses (calculated from 50 separate models). Effect is illustrated as partial dependence plots of each predictor with all other predictors held constant. Change in predictor is displayed on the *x*-axis and change in TN is displayed on the *y*-axis.

upper PLs rather than the single line that would be produced by distribution-based methods. For visual clarity, we plotted the envelopes containing individual upper PLs of training sites instead of the cloud of individual upper PLs themselves. Both methods identified identical numbers of training and validation sites as greater than their upper PL (Table 4). Prediction interval coverage probabilities (PICPs; the probability that all observed values fit within their prediction limits) calculated from validation data indicated that 90% and 94% of predictions were within the prediction limits for TP and TN, respectively, for both methods. Ideally, the PICP would equal the selected prediction limit of 95%. The TN model identified approximately the correct number of sites as above the upper PL, but upper PLs for the TP model were conservative, identifying more sites above the limit than expected.

SEE and UNEEC identified the same number of sites as having concentrations greater than the upper PL, but the specific sites identified as being over their PL varied between methods. For predicted high concentrations, upper PLs produced by the UNEEC method were larger than upper PLs produced by the SEE method, and the reverse was true for smaller predicted concentrations. This pattern occurred because of heteroscedasticity in model errors (seen in Fig. 4A, B), whereby larger predictions were made

TABLE 2. Assessment of random forest (RF) model performance and comparison with predictions of the Smith et al. (2003) model (r^2 = squared Pearson correlation coefficient) between observations and associated model predictions. The equivalent intercept is the percentage of 10,000 bootstrap simulations falling within the region of equivalence ($Eq_0 = \hat{Y} \pm 25\%$) for the intercept = 0. The equivalent slope is the percentage of 10,000 bootstrap simulations falling within the region of equivalence ($Eq_1 = m \pm 25\%$) for the slope = 1. Predictions made for training data are for out-of-bag data (i.e., not used in individual model creation). Tng = training data, Val = validation data, NSE = Nash–Sutcliffe Model Efficiency, RMSE = root mean square error, TP = total P, TN = total N.

Nutrient	Model	Data	n	r^2	NSE	RMSE	Equivalent intercept	Equivalent slope
TP	RF	Tng	752	0.40	0.40	16.2	100.0	100.0
		Val	40	0.46	0.43	20.5	83.8	22.2
	Smith et al.	Val	40	0.04	−0.10	28.5	56.1	16.4
TN	RF	Tng	665	0.32	0.32	113.9	100.0	99.6
		Val	35	0.23	0.16	80.1	96.8	34.6
	Smith et al.	Val	35	0	−0.58	109.6	75.7	0.4

with larger errors. The SEE method applies the same error to all predictions, and therefore, does not account for heteroscedasticity in model errors.

Discussion

Model performance

Our results showed that spatial variation in natural background TP and TN concentrations can be accurately predicted from geographic data, albeit not as precisely as we would like. We consider our models to be accurate. The TN model showed no consistent bias, and the bias of the TP model was <2% of the range of natural variation in TP concentration among our sites. Model predictions are generally applicable across the Mountain and Xeric ecoregions in the western USA, as demonstrated by model performance at validation sites for both models. Geothermal inputs can greatly affect nutrient concentrations, so streams with significant geothermal inputs are the major exception to the generality of our predictions. Catchments larger than those we used to develop our models (i.e., >9000 km^2) also would be outside of the experience of the model. We developed both models, in part, from data collected in the Great Plains. However, these sites did not span the range of environments found in the Great Plains, so our models should not be applied generally to streams in the Great Plains. The concordance of the

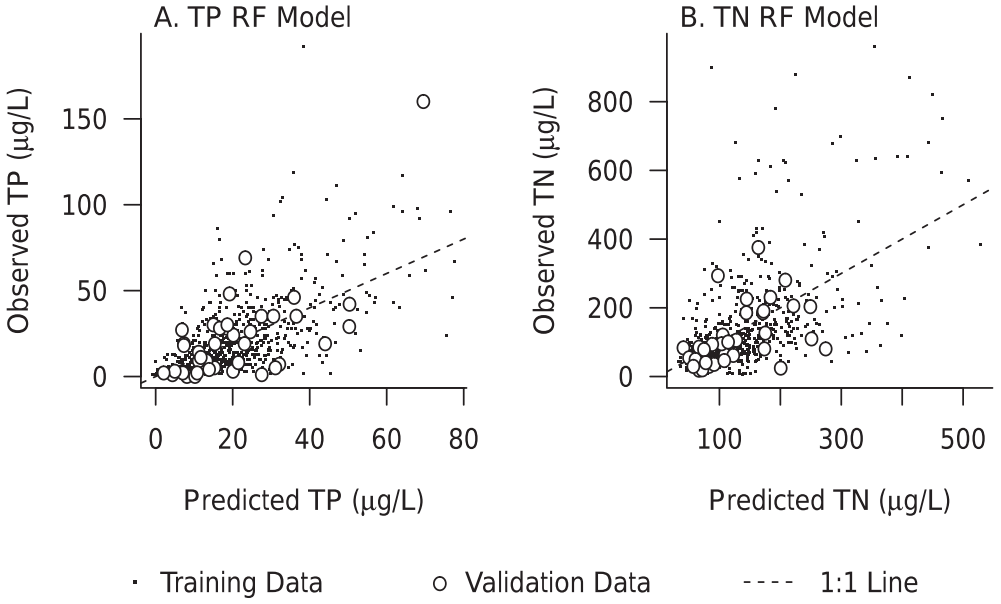


FIG. 4. Plots of observed vs predicted values of total P (TP) (A) and total N (TN) (B) for training and validation data for Random Forest (RF) models.

TABLE 3. Assessment of signal-to-noise (S:N) ratio. Var_{sites} = variance associated with sites, Var_{reps} = variance associated with replications, Max r^2 = highest possible r^2 value for a given S:N ratio calculated as S:N/(S:N + 1), TP = total P, TN = total N.

Data	Var_{sites} (signal)	Var_{reps} (noise)	S:N	Max r^2
TP	328	170	1.93	0.66
TN	11071	9311	1.19	0.54

observed relationships between predictors and nutrient concentrations with known mechanisms influencing TP and TN concentrations in streams further increases our confidence in the robustness of model predictions. The fact that the models accounted for most (59% for TP, 52% for TN) of the spatial variation in TP and TN concentrations indicates that the models were successful in capturing site-specific differences in reference conditions. We consider these models to be primarily spatial because the 1 or 2 predictors with

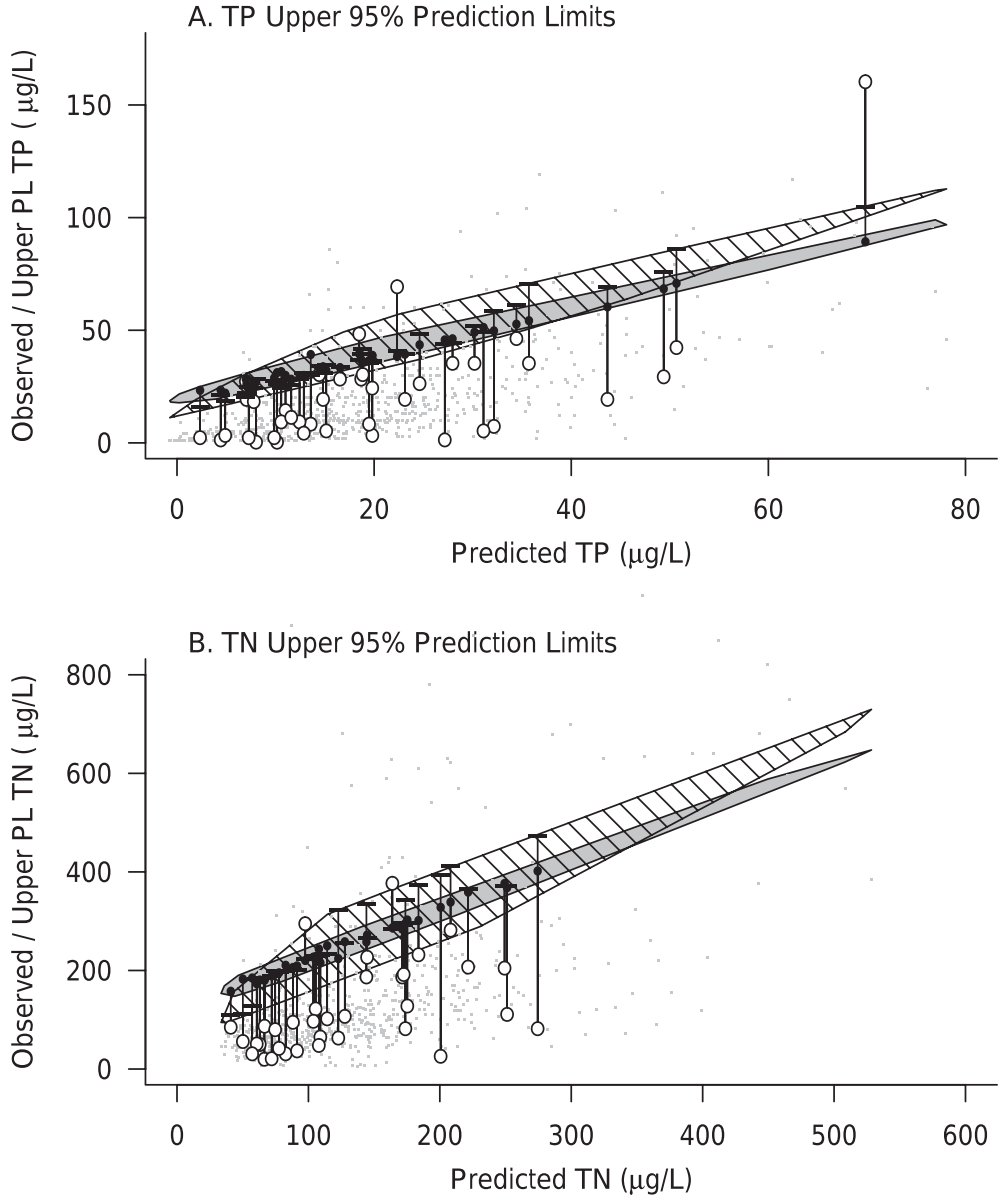


FIG. 5. Plots of observed vs predicted values and upper prediction limits (PLs) for total P (TP) (A) and total N (TN) (B) for training and validation data. Observations are plotted as grey dots (training data) or open circles (validation data). Regions containing upper PLs for training data are plotted as filled grey (Simple Empirical Error [SEE] method) or cross-hatch (UNcertainty Estimation based on Local Errors and Clustering [UNEEC] method; Shrestha and Solomatine 2008). Site-specific upper PLs for validation data are plotted as filled circles (SEE method) or bars (UNEEC method).

TABLE 4. Performance of upper prediction limits (PLs). For total P (TP) training data, $n = 752$, and for total N (TN) training data, $n = 665$. For TP validation data, $n = 40$, and for TN validation data, $n = 35$. UNEEC = UNcertainty Estimation based on Local Errors and Clustering method of Shrestha and Solomatine (2008), SEE = Simple Empirical Error, PICP = Prediction Interval Coverage Probability (Shrestha and Solomatine 2008).

	SEE method				UNEEC method			
	Training data		Validation data		Training data		Validation data	
Model	No. > upper PL	PICP	No. > upper PL	PICP	No. > upper PL	PICP	No. > upper PL	PICP
TP	68	91%	4	90%	68	91%	4	90%
TN	61	91%	2	94%	61	91%	2	94%

temporal components (i.e., previous year's precipitation in TP model, and day of year and prior 2 mo precipitation in TN model) were of only moderate importance in either model.

Model predictions based on measures of continuously varying environmental factors also clearly outperformed the Smith et al. (2003) model predictions. Comparing the performance of our models predicting baseflow concentrations with the Smith et al. models that predict annual flow-weighted concentrations is not ideal. However, models like those of Smith et al. created from means instead of individual samples could potentially yield better predictions because they should minimize the effects of measurement and temporal noise on model parameterization. The limited ability of the Smith et al. model to predict the baseflow-sample concentrations typically used in monitoring reveals a limitation inherent in models developed from mean sample concentrations. As also noted by Smith et al., their models were limited by their reliance on ecoregions for controlling spatial variation in nutrient sources. More recent applications of the SPARROW model (Garcia et al. 2011, Wise and Johnson 2011) account more directly for variation in natural sources of nutrients, but they predict only annual yields, which makes them difficult to use in monitoring. These models are probably better suited to predicting nutrient pollution levels and sources than background conditions because most applications of the SPARROW model are dominated by predictors related to anthropogenic sources.

Predictors

Most of the relationships between environmental factors and nutrient concentrations matched expectations based on previous studies, but relationships between nutrient concentrations and relative humidity, Ca deposition, EVI, precipitation, and SO_4^{2-} deposition were not as clearly related to known mechanisms. Increasing TP concentrations with decreasing humidity could be caused by evaporative concentration (Reddy et al. 1999). However, we see no reason to expect that atmospheric Ca deposition is directly linked to TP. Instead, the NADP measure of wet Ca deposition probably is correlated with dust deposition (Brahney 2012), and this variable may be acting as a surrogate for the deposition of P in dust (Reynolds et al. 2001). We expected TP and TN concentrations to decrease with increasing EVI because of increasing nutrient retention with increasing vegetation cover. However, this pattern occurred only in areas with lower EVI values associated with grasslands and scrub, and the opposite pattern occurred in areas with higher EVI values associated with forests (i.e., nutrient concentrations increased with increasing EVI). These increasing nutrient concentrations in forested areas might be attributable to lower nutrient retention by mature forest (Vitousek and Reiners 1975), built up litter fall from decades of fire suppression acting as a source of nutrients (Miller et al. 2005), or decreased microbial biomass resulting in lower P retention (Chen et al. 2003). More vegetation also could lead to increased rock weathering (as seen by Olson and Hawkins 2012 for other elements), which would release additional P.

The relationship between TN concentrations and precipitation also differed in direction of effect among environments. TN concentrations decreased with additional precipitation in xeric areas but increased with additional precipitation in mesic ones. TN concentrations have been observed to be positively correlated with precipitation in mesic areas (e.g., Hill 1986, Vanderbilt et al. 2003) and negatively correlated with precipitation in xeric areas (e.g., Lewis and Grant 1979, Alvarez-Cobelas et al. 2010), but these 2 patterns have not been observed in the same data set. Increasing precipitation in mesic areas can lead to higher TN concentrations resulting from increased N fixation in wet soils (Cleveland et al. 1999), litter decomposition (Lewis et al. 1999), and flushing caused by greater stream/hill slope connectivity

(Kane et al. 2008). Howarth et al. (2006) proposed that increased precipitation results in shorter water residence times that limit the amount of contact between runoff and denitrifying organisms in the stream bed. We suspect the negative relationship we observed between precipitation and TN concentrations in xeric areas is caused by water-dependent plant uptake. Greater precipitation in xeric areas may also create more anoxic zones in soils and, thus, increase denitrification (Bollmann and Conrad 1998). The relationship that is the least interpretable was the positive association between TN and atmospheric SO_4^{2-} deposition. This relationship is similar to the relationship seen by Cai et al. (2011) between stream NO_3^- and atmospheric SO_4^{2-} deposition in streams in Great Smoky Mountains National Park. SO_4^{2-} deposition could have a direct effect on stream TN by suppressing plant growth and, hence, N uptake, but we think it more likely that SO_4^{2-} deposition is a surrogate for another process or N source, such as dry deposition.

Volcanic rocks are a known source of P, but we were surprised that they were a more important predictor of stream TP than % rock P. During model development, we created models without % volcanic lithology as a predictor to assess its importance relative to % rock P. That model performed nearly as well as our TP model with % volcanic lithology (r^2 = 0.37 vs 0.40), and % rock P was the most important predictor, indicating that most, but not all, of the explanatory power of volcanic rocks is related to their P content. We attribute the remaining explanatory power of volcanic rocks to their relatively young age and faster weathering relative to other rock types (Gislason et al. 1996) in the western USA. Rapid weathering is especially true of recently active (within 1000–3000 y) basalt flows in the Gila Mountains/Mogollon Rim Ecoregion and may explain why streams in this region have average TP concentration >2× that of streams in the rest of our study area (48 μg/L vs 18 μg/L).

Several environmental factors associated with nutrient concentrations in other studies were not selected as predictors in our models. Rock N and dry N deposition both can be sources of N (Holloway and Dahlgren 2002, Fenn et al. 2003) that increase TN concentrations in streams and lakes. Rock N content was positively related to stream TN in our data as observed elsewhere (Williard et al. 2005, Gardner and McGlynn 2009), a result indicating that rock N is a source. However, this relationship was weak, and including it as a predictor did not improve model fit. Rock N may act as a significant source of stream TN only in specific circumstances where rock N content is high and readily weathered (e.g., Gardner and McGlynn 2009), such as in carbonaceous or oil shales. We also included estimates of dry N deposition derived from the CMAQ model in the TN model, but including these estimates slightly decreased model performance compared with models that included only wet N deposition (i.e., NADP data). This decrease in model performance with inclusion of dry N deposition estimates does not imply that dry deposition is not influencing stream TN, but rather that any potential model improvement associated with the inclusion of dry deposition was swamped by errors in deposition estimates. CMAQ dry-deposition estimates are based on emissions data instead of measured deposition as in the NADP data. Errors in deposition estimates could be caused by inaccurate emissions data, errors in the model estimating the distribution and amount of deposition, or both.

Factors associated with downstream nutrient losses and nutrient colimitation, both of which could potentially modify the amount of nutrients exported from catchments, also were not included in our models. Including catchment area, which is related to travel time and stream size and is associated with nutrient loss (Prairie and Kalff 1986, Smith et al. 2003), in our models decreased performance of both the TP and TN models. The lack of a relationship with catchment area in our study area probably occurred for several reasons. First, previous estimates of in-stream loss rates are mostly from agricultural catchments (e.g., Alexander et al. 2000), which have larger loss rates than reference catchments (Prairie and Kalff 1986, Mulholland et al. 2008). Greater uptake in streams flowing through agricultural catchments is probably caused by their higher nutrient concentrations, despite their lower uptake efficiencies (Mulholland et al. 2008). Second, although NH_4^+ uptake is positively related to stream size, the relationship between NO_3^- uptake and stream size is much noisier (Tank et al. 2008). The noisy NO_3^-–stream size relationship may obscure any effect that uptake of NH_4^+ by algae might have on TN concentrations because NO_3^- concentrations are much higher than NH_4^+ concentrations. Third, surrogates for denitrification (i.e., groundwater index) or streambed P adsorption or precipitation (i.e., Ca availability or channel slope) might have been more strongly associated with N and P removal because they are more direct surrogates of nutrient sinks than stream size. We also examined the possibility that P and N might be colimiting in streams as they are in lakes (Dodds et al. 2002). If N and P are colimiting, we would expect concentrations of one to be associated with concentrations of the other. For example, a

P-limited system would have lower N uptake and higher N export (and TN concentrations) at low P than at high P because of stoichiometric constraints on a stream's ability to use excess N. We assessed whether potential interactions between TP and TN improved predictions of each nutrient by including each nutrient as a predictor of the other. TP (either measured or predicted) had no effect on the performance of the TN model, but including measured TN slightly improved the r^2 of the TP model (0.40 to 0.42). However, we elected not to include TN as a predictor in the final TP model because the use of predicted TN did not improve the models and including measured TN as a predictor would prevent the application of these models to unmeasured locations.

Model shortcomings and possible improvements

Although the models made unbiased predictions of stream TP and TN concentrations in the western USA, these predictions could be potentially improved by addressing 2 model shortcomings. The 1st shortcoming of our models is their reliance on some predictors that can be altered by land use, which potentially could bias predictions of nutrient concentrations expected under natural conditions at altered sites. Vegetation predictors (e.g., EVI and % evergreen) may be especially problematic in this regard, but landuse alteration could also alter soil properties (bulk density and SOC). These predictors could simply be dropped from the models because they had relatively low importance, but a better approach would be to replace these predictors with estimates of potential vegetation (e.g., LANDFIRE Biophysical Settings Layer) or predicted natural soil properties (e.g., Malone et al. 2011). We did not pursue these options because it was not clear a priori which vegetation and soil attributes would be important.

A 2nd shortcoming of our models is their relatively poor precision. The effect of model imprecision is to increase upper PL, making criteria based upon these upper limits less protective than they would be if models were more precise. We attribute most of the poor model precision to temporal and measurement variation in grab-sample concentrations that was unaccounted for by our models. A comparison of the variation explained by our models with that potentially associated with spatial differences among streams indicates that most unexplained variation was some combination of this temporal and measurement error. Much of the unexplained temporal variation probably was associated with seasonal and yearly differences in runoff, flushing, freezing, or snowmelt. As models that characterize natural runoff and hydrologic regimes become available (e.g., Li et al. 2010), temporally and spatially explicit predictions of flow should enable better nutrient predictions (Helton et al. 2011). Also, some of the unexplained variation in nutrient concentrations may be the result of differences over time or between agencies in methods used to measure nutrient concentrations. TN measurements made before 1999 were almost 4× higher on average than measurements made after 1999, resulting in a positive relationship between year of sample and TN model residuals. This decrease in measured TN concentrations might be partially a result of the change from the Kjeldahl digestion method to persulfate oxidation and colorimetry method that occurred around this time. Patton and Kryskalla (2003) analyzed samples with both methods and observed that TN values obtained with persulfate oxidation and colorimetry were, on average, 15% lower than concentrations obtained with the Kjeldahl digestion method. Model performance probably could be improved by limiting data to observations measured with a single method or by adjusting concentrations to account for the method used (if that information is known). We chose to retain these earlier samples in our data to maximize the number of environments represented in our model, but recommend that future work be based on TN estimates derived from a single method. Developing models based on long-term average concentrations or loads should eliminate much of the residual error associated with temporal variation in grab-sample concentrations. However, using long-term averages to establish criteria for all of the streams that need to be assessed is not practical because of costs associated with such long-term measurements. A better approach would be to focus on predicting temporal variation in the nutrient concentrations observed from grab samples. Models that could predict both spatial and temporal variation would provide a better basis for establishing criteria and could provide potentially important ecological information on the location and timing of natural nutrient fluctuations that influence primary producers (e.g., Butzler and Chase 2009).

Much of the remaining unexplained spatial variation is probably associated with some combination of natural and anthropogenic factors not included in our models. Natural factors that we did not consider include inputs from migrating fish (either excreted or from carcasses), the effect of flow modification by beaver dams, variation in uptake with spatial or temporal changes in stream metabolism, and natural disturbances that affect catchment or riparian vegetation (e.g., Houlton et al. 2003, Eshleman et al. 2004). MODIS-derived EVI could be used to detect vegeta-

tion disturbances, but model development and application would then be restricted to the last 10 y, the period for which MODIS observations are available. Development of models of stream gross primary production and respiration (e.g., Bernot et al. 2010) would allow us to incorporate these metabolic factors that control nutrient uptake and denitrification rates (Mulholland et al. 2008). Potential anthropogenic sources of unexplained spatial variation include either historical (e.g., logging) or highly localized land use (e.g., cabins with septic systems near creeks), that were not caught by our screening. Dry N deposition and nutrient inputs delivered by dust are other potentially important anthropogenic sources (Ballantyne et al. 2011). Accounting for these inputs from national data sets like the NADP should be possible when our ability to measure or predict dry N deposition and dust improves.

Developing nutrient criteria

Both the SEE and UNEEC methods appear suitable for establishing upper PLs. PLs produced by both methods were conservative and found 1 to 5% more sites above their PL than expected from the chosen prediction interval (e.g., PICPs were 1 to 5% $<$ the chosen prediction interval of 95%). However, complete agreement may be difficult to achieve given that other applications of the UNEEC method resulted in PICPs that deviated from desired prediction levels by 4 to 9% (Solomatine and Shrestha 2009, Malone et al. 2011). The UNEEC method better accounted for data heteroscedasticity, but this modest improvement required a much more complicated approach that might limit understanding of the method by managers and stakeholders. The UNEEC method also assumes that prediction error is different under the different natural environmental conditions identified in the clustering step (Shrestha and Solomatine 2008). This assumption may be reasonable, but it has not been tested rigorously. Choice of method will involve a tradeoff between the potential to account for heteroscedasticity in prediction errors and ease in understanding how criteria are identified.

Quantifying prediction uncertainty allows regulators to address uncertainty explicitly when developing a criterion, an aspect of criteria-setting not often considered. We used an upper 95% PL in our example application, but the actual PL selection for use as a water-quality criterion should balance the likelihoods of under- and overprotection. PLs set too low increase the risk of overprotecting that stream and incurring unnecessary economic costs. PLs set too high increase the risk of environmental degradation and eutrophication. One approach to this tradeoff would be to adapt the tiered approach often used in setting biologic criteria (sensu Yoder and Rankin 1999). In this approach, 2 thresholds are set and sites are categorized as either ''meets reference'', ''needs additional monitoring'', or ''impaired''. The prediction and 95% PL could be used as thresholds, with values above the prediction triggering close monitoring and values above the 95% PL triggering immediate regulatory action.

Concluding remarks

Model-derived, site-specific criteria should better account for natural variation in nutrient concentrations than do regional criteria based on average regional conditions. As seen in other studies, observed nutrient concentrations for minimally altered reference sites varied over an order of magnitude within ecoregions (Fig. 6A–D). A comparison of this variation with proposed regional criteria (horizontal lines in Fig. 6A–D) highlights the difficulty of establishing a single criterion protective of most streams without overprotecting some significant minority of streams. For example, the criteria proposed by Herlihy and Sifneos (2008) and Smith et al. (2003) for TP in nutrient ecoregion II (Western Forested Mountains, Fig. 6A) would protect most sites, but would be overprotective of the 25% of sites with naturally high TP concentrations. The site-specific criteria identified for TP in this ecoregion by our approach are generally higher than these regional criteria, but avoid being overprotective. Also, in ~15% of cases, the site-specific criteria would be more protective than the regional criteria. This pattern of model-based upper PLs that are higher than the Herlihy and Sifneos (2008) regional criteria also occurred for TN in nutrient ecoregion II (Fig. 6C). In nutrient ecoregion III (Xeric West), our site-specific criteria generally were higher than the Smith et al. (2003) regional criteria for TP and TN (Fig. 6B, D). However, our PL-based site-specific criteria generally were lower than criteria developed from models by Dodds and Oakes (2004). The higher expected nutrient concentrations identified by Dodds and Oakes could have resulted from prediction error that occurs when effects of land use are not fully captured in landuse–nutrient models. Hill et al. (2013) noted that stream temperature models developed from only reference-site data predicted lower temperatures than did models built from data collected at both reference and nonreference sites that statistically controlled for the effects of land use. In some cases, model-based upper PLs agreed on average with proposed regional criteria

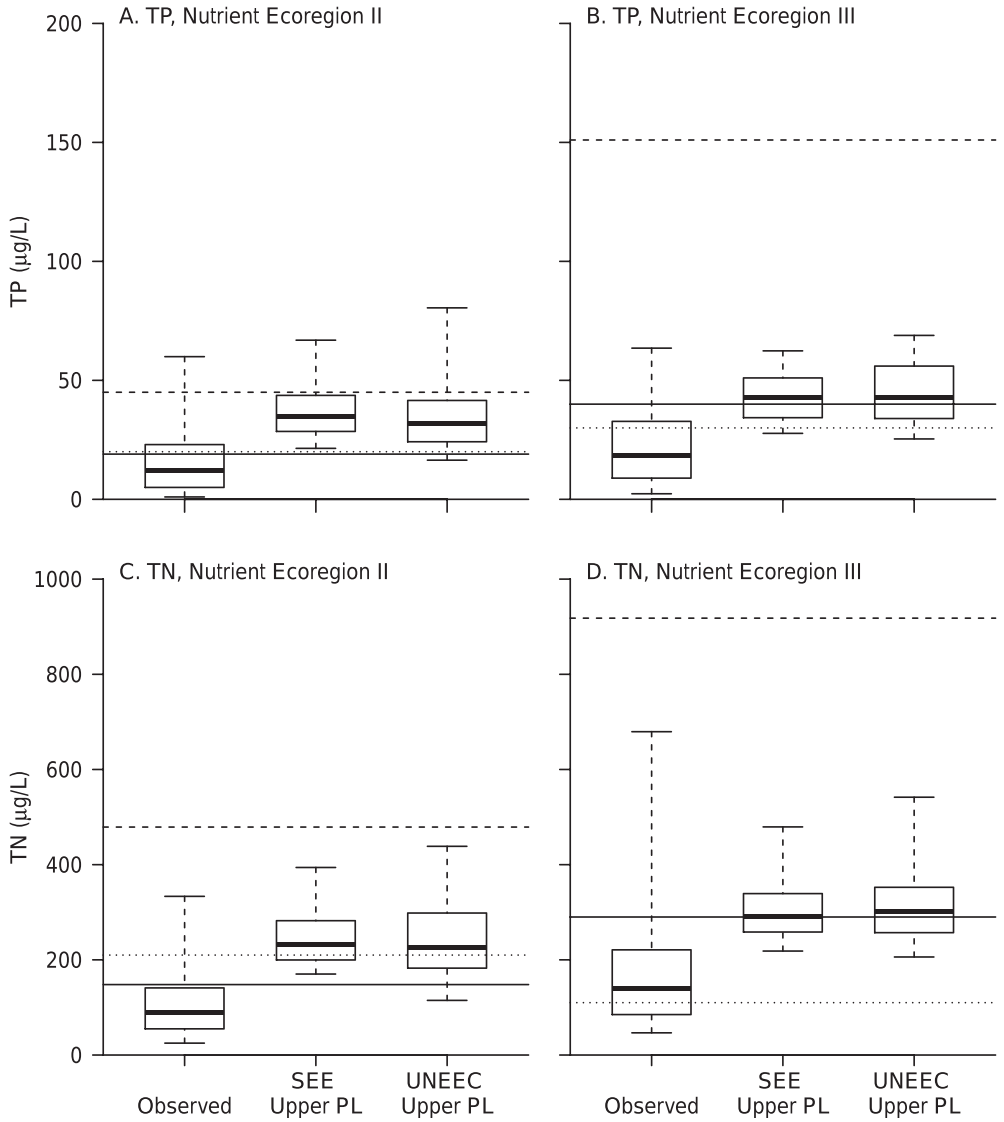


FIG. 6. Comparison of observed concentrations and upper prediction limits (PLs) for total P (TP) in Nutrient Ecoregion II Western Forested Mountains (A) and III Xeric West (B) and total N (TN) in Nutrient Ecoregions II (C) and III (D) with regional criteria from Herlihy and Sifneos (2008; solid lines), Dodds and Oakes (2004; dashed lines), and Smith et al. (2003; dotted lines). In all 4 cases, significant variation occurs within each region making any criterion identified over- or underprotective in many instances. Site-specific criteria based on upper PLs, although often higher than the regional criteria, better account for this observed variation. Lines in boxes are medians, box ends are quartiles, whiskers show 5th and 95th percentiles. SEE = Simple Empirical Error method, UNEEC = UNcertainty Estimation based on Local Errors and Clustering [UNEEC] method (Shrestha and Solomatine 2008).

(i.e., the Herlihy and Sifneos criterion in Fig. 6B, D or the Smith et al. criterion in Fig. 6C), but use of site-specific criteria would result in lower thresholds in ~½ of the cases and higher thresholds in the other ½.

The process of developing and applying site-specific criteria is more complex than the process of developing and applying regional criteria, but site-specific criteria are potentially more effective than regional criteria because they better account for natural variation. Sites with natural nutrient concentrations far below a regional criterion are underprotected and could potentially change trophic state while still meeting the criterion. Regional criteria applied to sites with naturally high concentrations may be challenged in court as too restrictive resulting in delay or prevention of implementation of the criterion. Given the complex processes that cause streams to differ in their natural nutrient concentrations, we think that setting nutrient

criteria based on regional classifications or typologies will be less effective than setting site-specific criteria (Hawkins et al. 2010).

Establishing meaningful nutrient criteria for individual streams is challenging but necessary for development and application of scientifically defensible and ecologically meaningful water-quality standards. Model-based, site-specific criteria will protect streams with naturally low nutrient concentrations from eutrophication better than regional criteria that are based, in part, on data from streams with naturally high concentrations. Conversely, streams with naturally higher nutrient concentrations should not be held to a standard that is impossible to achieve. Making site-specific predictions across large regions might appear challenging, but models based on readily available geographic predictors can now be developed easily and applied within a GIS framework to produce spatially explicit maps of expected nutrient conditions. Similar site-specific predictions have been made of streambed-surface grain sizes across France (Snelder et al. 2011). As additional data describing the spatial and temporal factors affecting nutrient concentrations become available, models can be improved to set nutrient criteria that are ever more reliable and protective.

Acknowledgements

This research was supported by grants R-828637-01 and R-830594-01 from the National Center for Environmental Research (NCER) Science to Achieve Results (STAR) Program of the US EPA. We thank Michelle Baker and Helga Van Miegroet for guidance and suggestions, Ryan Hill for GIS support, and John Van Sickle for his input on developing prediction intervals. Constructive suggestions for improving the manuscript were provided by Matt Baker, Michelle Baker, Peter Kolesar, Robin Jones, Helga Van Miegroet, John Van Sickle, Jacob Vander Laan, Associate Editor Lester Yuan, and 2 anonymous referees.

Literature Cited

ALEXANDER, R. B., R. A. SMITH, AND G. E. SCHWARZ. 2000. Effect of stream channel size on the delivery of nitrogen to the Gulf of Mexico. Nature 403:758–761.

ALVAREZ-COBELAS, M., R. SÁNCHEZ-ANDRÉS, S. SÁNCHEZ-CARRILLO, AND D. G. ANGELER. 2010. Nutrient contents and export from streams in semiarid catchments of central Spain. Journal of Arid Environments 74:933–945.

BAKER, M. E., M. J. WILEY, M. L. CARLSON, AND P. W. SEELBACH. 2003. A GIS model of subsurface water potential for aquatic resource inventory, assessment, and environmental management. Environmental Management 32: 706–719.

BALLANTYNE, A. P., J. BRAHNEY, D. FERNANDEZ, C. L. LAWRENCE, J. SAROS, AND J. C. NEFF. 2011. Biogeochemical response of alpine lakes to a recent increase in dust deposition in the Southwestern, US. Biogeosciences 8:2689–2706.

BERNOT, M. J., D. J. SOBOTA, R. O. HALL, P. J. MULHOLLAND, W. K. DODDS, J. R. WEBSTER, J. L. TANK, L. R. ASHKENAS, L. W. COOPER, C. N. DAHM, S. V. GREGORY, N. B. GRIMM, S. K. HAMILTON, S. L. JOHNSON, W. H. MCDOWELL, J. L. MEYER, B. PETERSON, G. C. POOLE, H. M. VALETT, C. ARANGO, J. J. BEAULIEU, A. J. BURGIN, C. A. CRENSHAW, A. M. HELTON, L. JOHNSON, J. MERRIAM, B. R. NIEDERLEHNER, J. M. O'BRIEN, J. D. POTTER, R. W. SHEIBLEY, S. M. THOMAS, AND K. WILSON. 2010. Inter-regional comparison of land-use effects on stream metabolism. Freshwater Biology 55:1874–1890.

BHADHA, J. H., AND J. W. JAWITZ. 2010. Characterizing deep soils from an impacted subtropical isolated wetland: implications for phosphorus storage. Journal of Soils and Sediments 10:514–525.

BOLLMANN, A., AND R. CONRAD. 1998. Influence of O_2 availability on NO and N_2O release by nitrification and denitrification in soils. Global Change Biology 4: 387–396.

BRAHNEY, J. 2012. The biogeochemical response of alpine lakes to changes in nutrient and dust deposition. PhD Dissertation, University of Colorado, Boulder, Colorado.

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

BREIMAN, L., J. H. FRIEDMAN, R. A. OLSHEN, AND C. J. STONE. 1984. Classification and regression trees. Wadsworth International Group, Belmont, California.

BUTZLER, J. M., AND J. M. CHASE. 2009. The effects of variable nutrient additions on a pond mesocosm community. Hydrobiologia 617:65–73.

CAI, M., J. S. SCHWARTZ, R. B. ROBINSON, S. E. MOORE, AND M. A. KULP. 2011. Long-term annual and seasonal patterns of acidic deposition and stream water quality in a Great Smoky Mountains high-elevation watershed. Water, Air, and Soil Pollution 219:547–562.

CEC (COMMISSION FOR ENVIRONMENTAL COOPERATION). 2006. Ecological regions of North America: toward a common perspective. US Environmental Protection Agency, Washington, DC. (Available from: http://www.epa.gov/wed/pages/ecoregions/na_eco.htm)

CHEN, C. R., L. M. CONDRON, M. R. DAVIS, AND R. R. SHERLOCK. 2003. Seasonal changes in soil phosphorus and associated microbial properties under adjacent grassland and forest in New Zealand. Forest Ecology and Management 177:539–557.

CHERUVELIL, K. S., P. A. SORANNO, M. T. BREMIGAN, T. WAGNER, AND S. L. MARTIN. 2008. Grouping lakes for water quality assessment and monitoring: the roles of regionalization and spatial scale. Environmental Management 41: 425–440.

CHÉTELAT, J., AND F. R. PICK. 2001. Temporal variability of water chemistry in flowing waters of the northeastern United States: does river size matter? Journal of the North American Benthological Society 20:331–346.

CHINNAYAKANAHALLI, K. 2006. The multi-watershed delineation tool. Utah State University, Logan, Utah. (Available from: http://hydrology.usu.edu/mwdtool/)

CLEVELAND, C. C., A. R. TOWNSEND, D. S. SCHIMEL, H. FISHER, R. W. HOWARTH, L. O. HEDIN, S. S. PERAKIS, E. F. LATTY, J. C. VON FISCHER, A. ELSEROAD, AND M. F. WASSON. 1999. Global patterns of terrestrial biological nitrogen (N_2) fixation in natural ecosystems. Global Biogeochemical Cycles 13:623–645.

COMPTON, J. E., M. R. CHURCH, S. T. LARNED, AND W. E. HOGSETT. 2003. Nitrogen export from forested watersheds in the Oregon Coast Range: the role of N_2-fixing red alder. Ecosystems 6:773–785.

CROSS, A. F., AND W. H. SCHLESINGER. 2001. Biological and geochemical controls on phosphorus fractions in semiarid soils. Biogeochemistry 52:155–172.

CUTLER, D. R., T. C. EDWARDS, K. H. BEARD, A. CUTLER, AND K. T. HESS. 2007. Random forests for classification in ecology. Ecology 88:2783–2792.

DALY, C., R. P. NEILSON, AND D. L. PHILLIPS. 1994. A statistical topographic model for mapping climatological precipitation over mountainous terrain. Journal of Applied Meteorology 33:140–158.

D'ARCY, P., AND R. CARIGNAN. 1997. Influence of catchment topography on water chemistry in southeastern Quebec Shield lakes. Canadian Journal of Fisheries and Aquatic Sciences 54:2215–2227.

DILLON, P. J., AND W. B. KIRCHNER. 1975. Effects of geology and land-use on export of phosphorus from watersheds. Water Research 9:135–148.

DILLON, P. J., AND L. A. MOLOT. 1997. Effect of landscape form on export of dissolved organic carbon, iron, and phosphorus from forested stream catchments. Water Resources Research 33:2591–2600.

DODDS, W. K. 2007. Trophic state, eutrophication and nutrient criteria in streams. Trends in Ecology and Evolution 22:669–676.

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.

DODDS, W. K., V. H. SMITH, AND K. LOHMAN. 2002. Nitrogen and phosphorus relationships to benthic algal biomass in temperate streams. Canadian Journal of Fisheries and Aquatic Sciences 59:865–874.

ESHLEMAN, K. N., D. A. FISCUS, N. M. CASTRO, J. R. WEBB, AND A. T. HERLIHY. 2004. Regionalization of disturbance-induced nitrogen leakage from mid-Appalachian forests using a linear systems model. Hydrological Processes 18:2713–2725.

FENN, M. E., J. S. BARON, E. B. ALLEN, H. M. RUETH, K. R. NYDICK, L. GEISER, W. D. BOWMAN, J. O. SICKMAN, T. MEIXNER, D. W. JOHNSON, AND P. NEITLICH. 2003. Ecological effects of nitrogen deposition in the western United States. BioScience 53:404–420.

GARCIA, A. M., A. B. HOOS, AND S. TERZIOTTI. 2011. A regional modeling framework of phosphorus sources and transport in streams of the Southeastern United States. Journal of the American Water Resources Association 47:991–1010.

GARDNER, K. K., AND B. L. MCGLYNN. 2009. Seasonality in spatial variability and influence of land use/land cover and watershed characteristics on stream water nitrate concentrations in a developing watershed in the Rocky Mountain West. Water Resources Research 45:W08411.

GISLASON, S. R., S. ARNORSSON, AND H. ARMANNSSON. 1996. Chemical weathering of basalt in southwest Iceland: effects of runoff, age of rocks and vegetative/glacial cover. American Journal of Science 296:837–907.

GLOBAL SOIL DATA TASK GROUP. 2000. Global Gridded Surfaces of Selected Soil Characteristics (International Geosphere–Biosphere Programme - Data and Information System). Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee. (Available from: http://www.daac.ornl.gov)

GREEN, M. B., AND J. C. FINLAY. 2010. Patterns of hydrologic control over stream water total nitrogen to total phosphorus ratios. Biogeochemistry 99:15–30.

GRIMALDI, C., AND V. CHAPLOT. 2000. Nitrate depletion during within-stream transport: effects of exchange processes between streamwater, the hyporheic and riparian zones. Water, Air, and Soil Pollution 124:95–112.

HAWKINS, C. P., J. R. OLSON, AND R. A. HILL. 2010. The reference condition: predicting benchmarks for ecological and water-quality assessments. Journal of the North American Benthological Society 29:312–343.

HELTON, A. M., G. C. POOLE, J. L. MEYER, W. M. WOLLHEIM, B. J. PETERSON, P. J. MULHOLLAND, E. S. BERNHARDT, J. A. STANFORD, C. ARANGO, L. R. ASHKENAS, L. W. COOPER, W. K. DODDS, S. V. GREGORY, R. O. HALL, S. K. HAMILTON, S. L. JOHNSON, W. H. MCDOWELL, J. D. POTTER, J. L. TANK, S. M. THOMAS, H. M. VALETT, J. R. WEBSTER, AND L. ZEGLIN. 2011. Thinking outside the channel: modeling nitrogen cycling in networked river ecosystems. Frontiers in Ecology and the Environment 9:229–238.

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, A. R. 1986. Stream nitrate-N loads in relation to variations in annual and seasonal runoff regimes. Water Resources Bulletin 22:829–839.

HILL, B. H., F. H. MCCORMICK, B. C. HARVEY, S. L. JOHNSON, M. L. WARREN, AND C. M. ELONEN. 2010. Microbial enzyme activity, nutrient uptake and nutrient limitation in forested streams. Freshwater Biology 55:1005–1019.

HILL, R. A., C. P. HAWKINS, AND D. M. CARLISLE. 2013. Predicting thermal reference conditions for USA streams and rivers. Freshwater Science 32:39–55.

HOLLOWAY, J. M., AND R. A. DAHLGREN. 2002. Nitrogen in rock: occurrences and biogeochemical implications. Global Biogeochemical Cycles 16:1118. doi:10.1029/2002GB001862

HOMER, C., C. Q. HUANG, L. M. YANG, B. WYLIE, AND M. COAN. 2004. Development of a 2001 National Land-Cover Database for the United States. Photogrammetric Engineering and Remote Sensing 70:829–840.

HOULTON, B. Z., C. T. DRISCOLL, T. J. FAHEY, G. E. LIKENS, P. M. GROFFMAN, E. S. BERNHARDT, AND D. C. BUSO. 2003. Nitrogen dynamics in ice storm-damaged forest ecosystems: implications for nitrogen limitation theory. Ecosystems 6:431–443.

HOUSE, W. A. 2003. Geochemical cycling of phosphorus in rivers. Applied Geochemistry 18:739–748.

HOWARTH, R. W., D. P. SWANEY, E. W. BOYER, R. MARINO, N. JAWORSKI, AND C. GOODALE. 2006. The influence of climate on average nitrogen export from large watersheds in the Northeastern United States. Biogeochemistry 79: 163–186.

HUETE, A., K. DIDAN, T. MIURA, E. P. RODRIGUEZ, X. GAO, AND L. G. FERREIRA. 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sensing of Environment 83:195–213.

ICE, G., AND D. BINKLEY. 2003. Forest streamwater concentrations of nitrogen and phosphorus: a comparison with EPA's proposed water quality criteria. Journal of Forestry 101:21–28.

KANE, E. S., E. F. BETTS, A. J. BURGIN, H. M. CLILVERD, C. L. CRENSHAW, J. B. FELLMAN, I. H. MYERS-SMITH, J. A. O'DONNELL, D. J. SOBOTA, W. J. VAN VERSEVELD, AND J. B. JONES. 2008. Precipitation control over inorganic nitrogen import–export budgets across watersheds: a synthesis of long-term ecological research. Ecohydrology 1:105–117.

KAUFMANN, P. R., P. LEVINE, E. G. ROBISON, C. SEELIGER, AND D. V. PECK. 1999. Quantifying physical habitat in wadeable streams. EPA/620/R-99/003. Western Ecology Division, US Environmental Protection Agency, Corvallis, Oregon.

KIRSCHBAUM, M. U. F. 2000. Will changes in soil organic carbon act as a positive or negative feedback on global warming? Biogeochemistry 48:21–51.

KNOWLTON, M. F., AND J. R. JONES. 2006. Natural variability in lakes and reservoirs should be recognized in setting nutrient criteria. Lake and Reservoir Management 22: 161–166.

LANDFIRE. 2011a. LANDFIRE 1.0.5 Biophysical settings layer. US Geological Survey, Reston, Virginia. (Available from: http://landfire.cr.usgs.gov/viewer/)

LANDFIRE. 2011b. LANDFIRE 1.0.5 Vegetation dynamics model descriptions. US Geological Survey, Reston, Virginia. (Available from: http://www.landfire.gov/NationalProductDescriptions24.php)

LEMMA. 2011. Landscape, Ecology, Modeling, Mapping, and Analysis Project. Oregon State University, Corvallis, Oregon. (Available from: http://www.fsl.orst.edu/lemma/main.php?project=nwfp15&id=studyAreas)

LEWIS, JR., W. M., AND M. C. GRANT. 1979. Relationships between stream discharge and yield of dissolved substances from a Colorado mountain watershed. Soil Science 128:353–363.

LEWIS, JR., W. M., J. M. MELACK, W. H. MCDOWELL, M. MCCLAIN, AND J. E. RICHEY. 1999. Nitrogen yields from undisturbed watersheds in the Americas. Biogeochemistry 46:149–162.

LI, M., Q. SHAO, L. ZHANG, AND F. H. S. CHIEW. 2010. A new regionalization approach and its application to predict flow duration curve in ungauged basins. Journal of Hydrology 389:137–145.

LIAW, A., AND M. WIENER. 2009. Package randomForests. R Project for Statistical Computing, Vienna, Austria. (Available from: http://cran.r-project.org/web/packages/randomForest/index.html)

LIKENS, G. E., C. T. DRISCOLL, D. C. BUSO, M. J. MITCHELL, G. M. LOVETT, S. W. BAILEY, T. G. SICCAMA, W. A. REINERS, AND C. ALEWELL. 2002. The biogeochemistry of sulfur at Hubbard Brook. Biogeochemistry 60:235–316.

LOPEZ, E. S., I. PARDO, AND N. FELPETO. 2001. Seasonal differences in green leaf breakdown and nutrient content of deciduous and evergreen tree species and grass in a granitic headwater stream. Hydrobiologia 464:51–61.

LUDINGTON, S., B. C. MORING, R. J. MILLER, P. A. STONE, A. A. BOOKSTROM, D. R. BEDFORD, J. G. EVANS, G. A. HAXEL, C. J. NUTT, K. S. FLYN, AND M. J. HOPKIN. 2007. Preliminary integrated geologic map databases for the United States Western States: California, Nevada, Arizona, Washington, Oregon, Idaho, and Utah. U.S. Geological Survey Open-File Report 2005-1305. US Geological Survey, Reston, Virginia. (Available from: http://pubs.usgs.gov/of/2005/1305/)

MALONE, B. P., A. B. MCBRATNEY, AND B. MINASNY. 2011. Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes. Geoderma 160: 614–626.

MARCARELLI, A. M., AND W. A. WURTSBAUGH. 2006. Temperature and nutrient supply interact to control nitrogen fixation in oligotrophic streams: an experimental examination. Limnology and Oceanography 51:2278–2289.

MEINSHAUSEN, N. 2006. Quantile regression forests. Journal of Machine Learning Research 7:983–999.

MEYBECK, M. 1982. Carbon, nitrogen, and phosphorus transport by world rivers. American Journal of Science 282:401–450.

MILLER, W. W., D. W. JOHNSON, C. DENTON, P. S. J. VERBURG, G. L. DANA, AND R. F. WALKER. 2005. Inconspicuous nutrient laden surface runoff from mature forest Sierran watersheds. Water, Air, and Soil Pollution 163:3–17.

MULHOLLAND, P. J. 1992. Regulation of nutrient concentrations in a temperate forest stream: roles of upland, riparian, and instream processes. Limnology and Oceanography 37:1512–1526.

MULHOLLAND, P. J., A. M. HELTON, G. C. POOLE, R. O. HALL, S. K. HAMILTON, B. J. PETERSON, J. L. TANK, L. R. ASHKENAS, L. W. COOPER, C. N. DAHM, W. K. DODDS, S. E. G. FINDLAY, S. V. GREGORY, N. B. GRIMM, S. L. JOHNSON, W. H. MCDOWELL, J. L. MEYER, H. M. VALETT, J. R. WEBSTER, C. P. ARANGO, J. J. BEAULIEU, M. J. BERNOT, A. J. BURGIN, C. L. CRENSHAW, L. T. JOHNSON, B. R. NIEDERLEHNER, J. M. O'BRIEN, J. D. POTTER, R. W. SHEIBLEY, D. J. SOBOTA, AND S. M. THOMAS. 2008. Stream denitrification across biomes and its response to anthropogenic nitrate loading. Nature 452:202–205.

NRCS (NATURAL RESOURCES CONSERVATION SERVICE). 2011. U.S. General Soil Map (STATSGO2). Natural Resources

Conservation Service, Washington, DC. (Available from: http://soildatamart.nrcs.usda.gov)

OHMANN, J. L., M. J. GREGORY, AND T. A. SPIES. 2007. Influence of environment, disturbance, and ownership on forest vegetation of Coastal Oregon. Ecological Applications 17:18–33.

OLSON, J. R., AND C. P. HAWKINS. 2012. Predicting natural base-flow stream water chemistry in the western United States. Water Resources Research 48:WR011088.

PANAGOPOULOS, I., M. MIMIKOU, AND M. KAPETANAKI. 2007. Estimation of nitrogen and phosphorus losses to surface water and groundwater through the implementation of the SWAT model for Norwegian soils. Journal of Soils and Sediments 7:223–231.

PARK, J. H., M. J. MITCHELL, P. J. MCHALE, S. F. CHRISTOPHER, AND T. P. MEYERS. 2003. Impacts of changing climate and atmospheric deposition on N and S drainage losses from a forested watershed of the Adirondack Mountains, New York State. Global Change Biology 9:1602–1619.

PATTON, C. J., AND J. R. KRYSKALLA. 2003. Methods of analysis by the U.S. Geological Survey National Water Quality Laboratory. Evaluation of alkaline persulfate digestion as an alternative to Kjeldahl digestion for determination of total and dissolved nitrogen and phosphorus in water. Water-Resources Investigations Report 03-4174. US Geological Survey, Denver, Colorado.

PIÑEIRO, G., S. PERELMAN, J. P. GUERSCHMAN, AND J. M. PARUELO. 2008. How to evaluate models: observed vs. predicted or predicted vs. observed? Ecological Modelling 216: 316–322.

PRAIRIE, Y. T., AND J. KALFF. 1986. Effect of catchment size on phosphorus export. Water Resources Bulletin 22: 465–470.

RAMPE, J. J., R. D. JACKSON, AND M. R. SOMMERFELD. 1981. Physicochemistry of the upper Gila River watershed: I. San Francisco River and Clifton Hot Springs. Arizona-Nevada Academy of Science Journal 16:1–6.

REDDY, K. R., R. H. KADLEC, E. FLAIG, AND P. M. GALE. 1999. Phosphorus retention in streams and wetlands: a review. Critical Reviews in Environmental Science and Technology 29:83–146.

REYNOLDS, R., J. BELNAP, M. REHEIS, P. LAMOTHE, AND F. LUISZER. 2001. Aeolian dust in Colorado Plateau soils: nutrient inputs and recent change in source. Proceedings of the National Academy of Sciences of the United States of America 98:7123–7127.

ROBERTSON, D. M., D. A. SAAD, AND D. M. HEISEY. 2006. A regional classification scheme for estimating reference water quality in streams using land-use-adjusted spatial regression-tree analysis. Environmental Management 37:209–229.

ROBINSON, A. P., R. A. DUURSMA, AND J. D. MARSHALL. 2005. A regression-based equivalence test for model validation: shifting the burden of proof. Tree Physiology 25:903–913.

SCHWEDE, D. B., R. L. DENNIS, AND M. A. BITZ. 2009. The watershed deposition tool: a tool for incorporating atmospheric deposition in water-quality analyses. Journal of the American Water Resources Association 45: 973–985.

SHRESTHA, D. L., AND D. P. SOLOMATINE. 2008. Data-driven approaches for estimating uncertainty in rainfall-runoff modelling. International Journal of River Basin Management 6:109–122.

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 and Technology 37:3039–3047.

SNELDER, T. H., B. J. F. BIGGS, AND M. A. WEATHERHEAD. 2004. Nutrient concentration criteria and characterization of patterns in trophic state for rivers in heterogeneous landscapes. Journal of the American Water Resources Association 40:1–13.

SNELDER, T. H., N. LAMOUROUX, AND H. PELLA. 2011. Empirical modelling of large scale patterns in river bed surface grain size. Geomorphology 127:189–197.

SOLOMATINE, D. P., AND D. L. SHRESTHA. 2009. A novel method to estimate model uncertainty using machine learning techniques. Water Resources Research 45:W00B11.

STEVENSON, F. J., AND M. A. COLE. 1999. Cycles of soil: carbon, nitrogen, phosphorus, sulfur, micronutrients. 2nd edition. John Wiley and Sons, New York.

STODDARD, J. L., A. T. HERLIHY, D. V. PECK, R. M. HUGHES, T. R. WHITTIER, AND E. TARQUINIO. 2008. A process for creating multimetric indices for large-scale aquatic surveys. Journal of the North American Benthological Society 27:878–891.

STOESER, D. B., G. N. GREEN, L. C. MORATH, W. D. HERAN, A. B. WILSON, D. W. MOORE, AND B. S. VAN GOSEN. 2007. Preliminary integrated geologic map databases for the United States: central States: Montana, Wyoming, Colorado, New Mexico, North Dakota, South Dakota, Nebraska, Kansas, Oklahoma, Texas, Iowa, Missouri, Arkansas, and Louisiana. U.S. Geological Survey Open-File Report 2005-1351. US Geological Survey, Reston, Virginia. (Available from: http://pubs.usgs.gov/of/2005/1351/)

STROBL, C., A. L. BOULESTEIX, T. KNEIB, T. AUGUSTIN, AND A. ZEILEIS. 2008. Conditional variable importance for random forests. BMC Bioinformatics 9:307.

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.

TANK, J. L., E. J. ROSI-MARSHALL, M. A. BAKER, AND R. O. HALL. 2008. Are rivers just big streams? A pulse method to quantify nitrogen demand in a large river. Ecology 89:2935–2945.

TORBERT, H. A., AND C. W. WOOD. 1992. Effects of soil compaction and water-filled pore-space on soil microbial activity and N losses. Communications in Soil Science and Plant Analysis 23:1321–1331.

USEPA (US ENVIRONMENTAL PROTECTION AGENCY). 2000. Nutrient criteria technical guidance manual, rivers and streams. EPA-822-B-00-002. Office of Water, US Environmental Protection Agency, Washington, DC.

USEPA (US ENVIRONMENTAL PROTECTION AGENCY). 2011. Working in partnership with states to address phosphorus and nitrogen pollution through use of a framework for state nutrient reductions. Office of Water, US Environmental Protection Agency, Washington, DC.

USGS (US GEOLOGICAL SURVEY). 2006. National Hydrography Dataset (NHD) Medium Resolution. US Geological Survey, Reston, Virginia. (Available from: http://nhd.usgs.gov/data.html)

VANDERBILT, K. L., K. LAJTHA, AND F. J. SWANSON. 2003. Biogeochemistry of unpolluted forested watersheds in the Oregon Cascades: temporal patterns of precipitation and stream nitrogen fluxes. Biogeochemistry 62:87–117.

VAN MIEGROET, H., I. F. CREED, N. S. NICHOLAS, D. G. TARBOTON, K. L. WEBSTER, J. SHUBZDA, B. ROBINSON, J. SMOOT, D. W. JOHNSON, S. E. LINDBERG, G. LOVETT, S. NODVIN, AND S. MOORE. 2001. Is there synchronicity in nitrogen input and output fluxes at the Noland Divide Watershed, a small N-saturated forested catchment in the Great Smoky Mountains National Park? TheScientificWorld 1:480–492.

VITOUSEK, P. M., AND W. A. REINERS. 1975. Ecosystem succession and nutrient retention: a hypothesis. BioScience 25:376–381.

WALKER, T. W., AND J. K. SYERS. 1976. Fate of phosphorus during pedogenesis. Geoderma 15:1–19.

WANG, G. P., J. S. LIU, H. Y. ZHAO, J. D. WANG, AND J. B. YU. 2007. Phosphorus sorption by freeze–thaw treated wetland soils derived from a winter-cold zone (Sanjiang Plain, Northeast China). Geoderma 138:153–161.

WASHBURN, C. S. M., AND M. A. ARTHUR. 2003. Spatial variability in soil nutrient availability in an oak-pine forest: potential effects of tree species. Canadian Journal of Forest Research–Revue Canadienne de Recherche Forestière 33:2321–2330.

WILLIARD, K. W. J., D. R. DEWALLE, AND P. J. EDWARDS. 2005. Influence of bedrock geology and tree species composition on stream nitrate concentrations in mid-Appalachian forested watersheds. Water, Air, and Soil Pollution 160:55–76.

WISE, D. R., AND H. M. JOHNSON. 2011. Surface-water nutrient conditions and sources in the United States Pacific Northwest. Journal of the American Water Resources Association 47:1110–1135.

YODER, C. O., AND E. T. RANKIN. 1999. Biological criteria for water resource management. Pages 227–259 *in* P. C. Schulze (editor). Measures of environmental performance and ecosystem condition. National Academy Press, Washington, DC.

ZHANG, G., AND Y. LU. 2012. Bias-corrected random forests in regression. Journal of Applied Statistics 39:151–160.

Received: 23 July 2012
Accepted: 5 April 2013