The Annals of Applied Statistics
2020, Vol. 14, No. 4, 1651–1675
https://doi.org/10.1214/20-AOAS1371

DATA FUSION MODEL FOR SPECIATED NITROGEN TO IDENTIFY ENVIRONMENTAL DRIVERS AND IMPROVE ESTIMATION OF NITROGEN IN LAKES

BY ERIN M. SCHLIEP[1,*], SARAH M. COLLINS[2], SHIRLEY ROJAS-SALAZAR[1,†], NOAH R. LOTTIG[3,‡] AND EMILY H. STANLEY[3,§]

[1]*Department of Statistics, University of Missouri,* [*]*schliepe@missouri.edu;* [†]*srnw3@mail.missouri.edu*
[2]*Department of Zoology and Physiology, University of Wyoming, sarah.collins@uwyo.edu*
[3]*Center for Limnology, University of Wisconsin,* [‡]*nrlottig@wisc.edu;* [§]*ehstanley@wisc.edu*

Concentrations of nitrogen provide a critical metric for understanding ecosystem function and water quality in lakes. However, varying approaches for quantifying nitrogen concentrations may bias the comparison of water quality across lakes and regions. Different measurements of total nitrogen exist based on its composition (e.g., organic versus inorganic, dissolved versus particulate), which we refer to as nitrogen species. Fortunately, measurements of multiple nitrogen species are often collected and can, therefore, be leveraged together to inform our understanding of the controls on total nitrogen in lakes. We develop a multivariate hierarchical statistical model that fuses speciated nitrogen measurements, obtained across multiple methods of reporting, in order to improve our estimates of total nitrogen. The model accounts for lower detection limits and measurement error that vary across lake, species and observation. By modeling speciated nitrogen, as opposed to previous efforts that mostly consider only total nitrogen, we obtain more resolved inference with regard to differences in sources of nitrogen and their relationship with complex environmental drivers. We illustrate the inferential benefits of our model using speciated nitrogen data from the LAke GeOSpatial and temporal database (LAGOS).

1. Introduction. Water quality in freshwater ecosystems is often controlled by the availability of nutrients. Historically, most research focused on phosphorus as the foremost control on primary production in lakes (Schindler et al. (2008), Schindler (2012)), but recent studies have highlighted a critical role for nitrogen (N) in shaping ecosystem function and water quality (Paerl et al. (2016), Harpole et al. (2011)). Both the amount and the form in which N occurs is significant, as the inorganic forms of this nutrient (defined below) have been associated with the formation of toxin-forming algal blooms (Glibert et al. (2016), Gobler et al. (2016)). While it has become apparent that understanding the sources and cycling of N is imperative for characterizing and managing lake water quality, several factors have made this task challenging. First, the total quantity of nitrogen (total nitrogen or TN) in lake water can be broken down into several forms (e.g., organic vs. inorganic, dissolved vs. particulate, hereafter referred to as N "species"); each species may have different sources, and there are complex environmental processes (referred to as drivers) of the sources of and transformations between species (Wetzel (2001)). Second, different measurement methods with varying detection limits can be used to "observe" TN (Saunders et al. (2017), Stow et al. (2018)) which can lead to potential biases when compared across time and lakes. Finally, there is dramatic variation in nutrient levels among lakes (Read et al. (2015)) as well as lake and watershed characteristics across continental scales (Hill et al. (2018)), necessitating macroscale

Received September 2019; revised March 2020.
Key words and phrases. Bayesian hierarchical model, detection limits, LAGOS, multivariate, Markov chain Monte Carlo.

studies to develop a generalized understanding of controls on TN. We address these challenges by developing a multivariate hierarchical model that can be used to simultaneously evaluate environmental drivers of multiple species of nitrogen in lakes across continental scales and can accommodate complex data that include different measurement methods, data sources and detection limits.

Measurable nitrogen species in lakes include total nitrogen (TN), total kjeldahl nitrogen (TKN), ammonium (NH_4) and nitrate-nitrite (NO_2NO_3). Letting DOPN denote dissolved organic and particulate nitrogen, which is never measured directly, TN can be decomposed as

$$\text{TN} = \text{DOPN} + \text{NH}_4 + \text{NO}_2\text{NO}_3, \tag{1}$$

where

$$\text{TKN} = \text{DOPN} + \text{NH}_4. \tag{2}$$

The two common measurement methods for TN include: (i) directly measured TN data and (ii) TN calculated from the sum of TKN and NO_2NO_3. To our knowledge, neither the distinction between these methods nor the fusion of measurements of multiple subspecies of nitrogen have been accounted for in a statistical model-based framework.

A number of landscape characteristics (Table 1) can influence the delivery of TN from land to lake ecosystems, and these processes vary dramatically across the United States (Read et al. (2015), Collins et al. (2017)). Variables that are broadly structured over space, such as climate or agriculture, can influence the magnitude of N sources and transport through watersheds (Chen et al. (2015)). Local characteristics, such as lake depth, can also have a strong influence on nutrient concentrations in lakes at continental scales (Read et al. (2015)). Examples of these broad and local characteristics are shown spatially in Figure 12 of Appendix B. Collectively, we refer to these environmental characteristics as possible environmental drivers of nitrogen which are often treated as covariates in a statistical regression model.

The effects of environmental drivers might influence each N species differently. For example, row-crop agriculture is likely to have a strong relationship with NO_2NO_3 due to pervasive fertilizer use and susceptibility of soils to leach nitrate to receiving streams and lakes (Cameron, Di and Moir (2013), Dorioz and Ferhi (1994)). Whereas previous analyses with continental or subcontinental data have developed relationships between only total nitrogen and important environmental drivers (e.g., Knoll et al. (2015), Read et al. (2015), Collins et al. (2017)) or the relationships between the species but not the environment (e.g., Wu et al. (2017)), our model allows for both nitrogen species-specific relationships with these environmental drivers and possible dependence between species. Understanding potentially synergistic or contradictory relationships between each N species and environmental driver

TABLE 1
Environmental drivers of lake nutrients at the lake or regional scale

Region-specific	baseflow
	runoff
	atmospheric nitrogen deposition
	watershed land use (% row crop, % forest, % wetland)
Lake-specific	lake area
	maximum depth
	lake-to-watershed area ratio
	measures of lake connectivity

will give additional insight into our understanding of N dynamics and could be useful for management.

We develop the model using data from LAGOS-NE v.1.087.3,[1] a lake water quality database that assimilated and harmonized lake nutrient data from 87 agency, university, tribal and citizen monitoring programs and includes different methods for measurement and detection levels (Soranno et al. (2015a), Soranno et al. (2017), Soranno and Cheruvelil (2017)). We also use auxiliary data from the United States Geological Survey (USGS) Standard Reference Sample Project[2] and North-Temperate Lakes Long Term Ecological Research (LTER) program[3] to quantify the specific measurement errors for each species of nitrogen. Included in LAGOS-NE are N species data from several thousand lakes across a 17-state region in the Northeastern and Midwestern United States, providing a unique opportunity to examine drivers of different N species at a broad spatial extent with heterogeneous land use, climate and lake characteristics. A detailed description of the LAGOS-NE data used in this analysis as well as the auxiliary data is provided in Section 3.

The process of building LAGOS-NE included a QA/QC procedure intended to harmonize data from different sampling programs so, they can be compared over space and time (Soranno et al. (2015a)), but different methods of quantifying total nitrogen exist across sampling programs, including both directly measured and calculated TN. While an in-depth analysis showed that these methods generate comparable results (Stanley et al. (2019b)), sampling bias might exist. For the LAGOS-NE lakes used in this analysis, Figure 1 shows the median annual reported value of total N for the years 1980–2012 as well as the method used to report total N, either measured or calculated. The method of reporting varies across LAGOS-NE

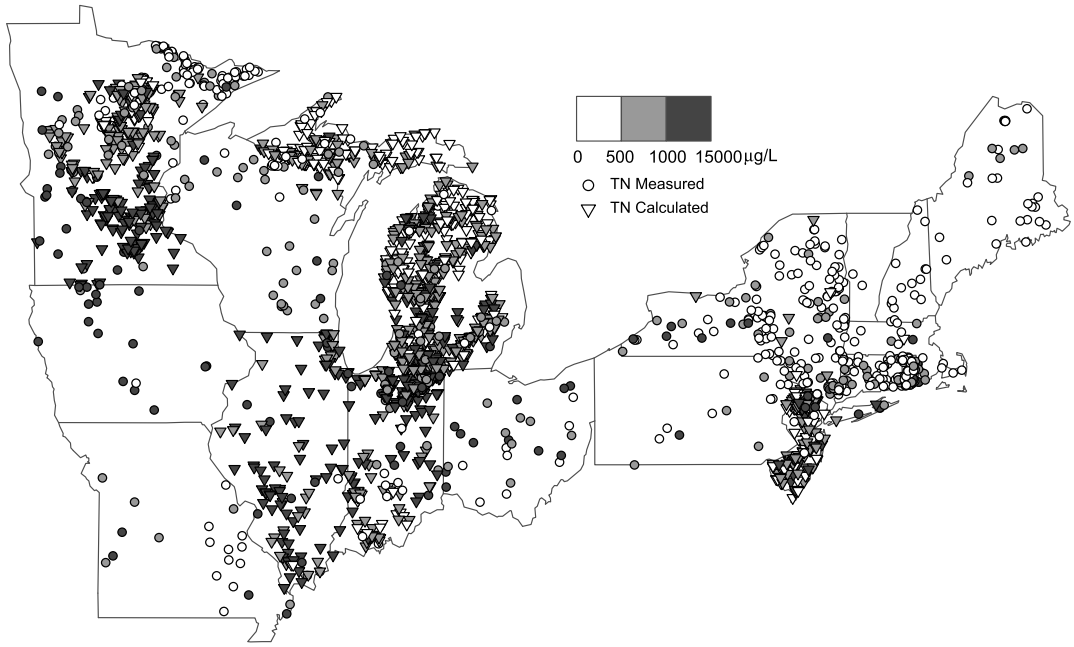

FIG. 1. *Median observations of TN across LAGOS-NE by method of reporting (measured or calculated) showing state and regional variation and clustering in both N concentration (µg/L) and method.*

[1]Data from LAGOS-NE are available at https://lagoslakes.org/products/data-products/. See the `LAGOSNE` R package to load and use the data (Stachelek and Oliver (2017)).

[2]Data from the USGS Standard Reference Sample Project are available at https://bqs.usgs.gov/srs/.

[3]Data from the North-Temperate Lakes Long Term Ecological Research program are available at https://lter.limnology.wisc.edu/.

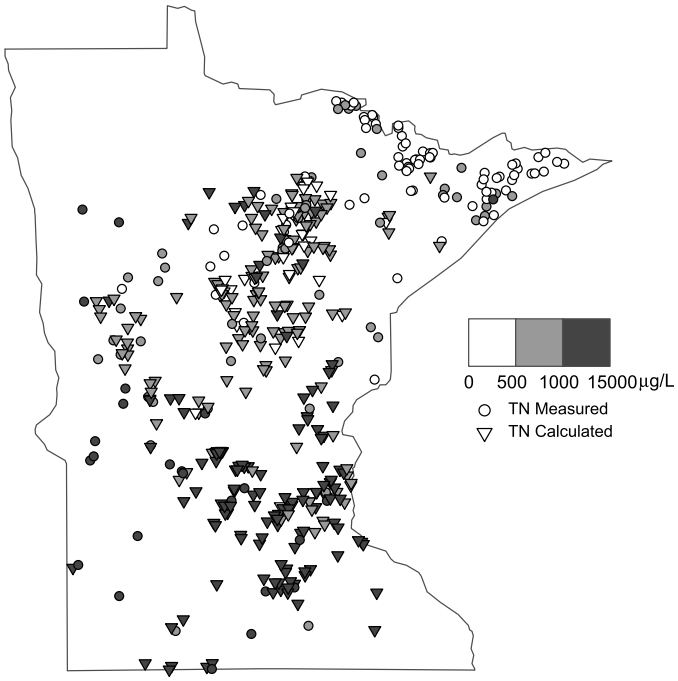

FIG. 2. *Median observations of TN by method of reporting (measured or calculated) for the state of Minnesota showing a north-south gradient in N concentration ($\mu g/L$).*

and has strong state and regional bias toward calculated TN in high-nutrient lakes in the upper Midwest and measured TN in low-nutrient lakes in the Northeast. Figure 2 provides an enhanced view of the nutrient data and reporting method for the state of Minnesota. Lakes in northern Minnesota tend to have low nitrogen concentrations compared to those in the south. Both methods of reporting are present in Minnesota; however, measured TN appears to be favored in the north, whereas the two methods are more equally favored in the central and southern portion of the state. Another common difference among sampling programs is that the lower limit of detection for nitrogen is highly variable or often unreported (Stow et al. (2018)). In addition, both the detection limit and measurement error for the different species of nitrogen (e.g., organic versus inorganic) can vary (Saunders et al. (2017)). Most inferential models developed to understand drivers of lake TN fail to include any information about detection limits for individual observations despite their potential importance (e.g., Oliver et al. (2017), Wagner and Schliep (2018)). Models that accommodate these technical differences among programs (i.e., measured vs. calculated, detection limits and measurement error) will enhance our understanding of the patterns and drivers of lake water quality but, as of yet, have not been widely used.

The challenge of modeling data from different sources is common in environmental and ecological applications where observations may be obtained from a combination of ground-level data (e.g., field data, monitoring stations) and remote-sensing data (e.g., satellites, LiDAR). In addition, these data sources might vary in terms of the type of data collected, such as species presence/absence data or abundance and the spatial and/or temporal resolution of the data product (e.g., point-level data, areal-unit data). To leverage the information provided from mismatched data sources, types and scales, data fusion approaches (e.g., Fuentes and Raftery (2005), Pacifici et al. (2017), Hilker et al. (2009)) and statistical downscalers (e.g., Guillas et al. (2008), Berrocal, Gelfand and Holland (2010)) have been proposed. Modifications and extensions of joint distribution models have also been developed that account for different data types (including censored data) as well as heterogeneous sampling efforts

(e.g., generalized joint attribute models, Clark et al. (2017)). The variability in data collection might exist due to variations in plot sizes, search times or gear type, each of which can greatly influence inference and prediction if not accounted for in the model.

Our work borrows ideas from Rundel et al. (2015), who propose a multivariate statistical model for air quality metrics using data from three ground monitoring station networks and gridded computer model data. The data from these four sources include various measurements of speciated and total particulate matter ($PM_{2.5}$). Their multilevel model specification using latent processes ensures that total PM is equal to the sum of its components (which, like in LAGOS, is not necessarily true in the observed data) and allows for variability in measurement error from the different sources and variable type. Their work directly models the spatial variability of $PM_{2.5}$ and its speciated components using spatially dependent processes, as opposed to modeling the variability by regressing on auxiliary explanatory variables. Importantly, they are not tasked with incorporating lower detection limits in their model nor interested in understanding the variability in environmental drivers of the variables of interest.

The remainder of this article is organized as follows. In Section 2 we develop the statistical data-fusion model for total nitrogen that addresses the challenges of multiple methods for obtaining TN and speciated nitrogen observations as well as the variability in lower detection limits and measurement error. For comparison, an analogous univariate model for TN is also presented. Model inference is obtained in a Bayesian framework; details of which are given in Section 2. The speciated nitrogen model is applied to the LAGOS-NE data from the 17 state region of the northeast United States in Section 3. Inference from the multivariate model is presented and compared to that from the univariate model. Section 4 concludes with a discussion and directions for further research.

2. Speciated nitrogen model specification. The four measurable quantities of speciated nitrogen in lakes include TN, TKN, NH_4 and NO_2NO_3. Due to the variability in observers making the measurements of speciated nitrogen at each lake, all four measurements are rarely reported simultaneously. As shown in (1) and (2), each measurable quantity can be written as a linear combination of the three distinct species: DOPN, NH_4, and NO_2NO_3. From a modeling perspective, by decomposing TN into these three nonnegative components, we can ensure the proper inequalities between the variables are held. In particular, $TKN > NH_4$, $TN > TKN$ and $TN > NO_2NO_3$. Due to measurement error, which includes human error in reporting, these inequalities are often violated in the reported speciated nitrogen data in LAGOS. For reference, in the data used in our analysis, 1%, 6% and 15% of lakes have observations that violate the three inequalities, respectively, and no identifiable patterns of these violations were detected.

The purpose of this analysis is to build a model that leverages the information across all measurements of speciated nitrogen in order to: (i) better estimate total nitrogen and (ii) make more resolved inference with regard to the environmental drivers of speciated nitrogen. To this end, we propose a multivariate hierarchical statistical model for the three species: DOPN, NH_4 and NO_2NO_3. The model will allow for dependence between speciated nitrogen concentrations and account for the variability in method of reported TN (measured vs. calculated) across the landscape and the associated error in observations. In addition, each of the measurable variables is subject to a detection limit which is lake specific. That is, for each variable and each lake there is a lower bound on the detectable value for that variable. This detection information will be incorporated into the multivariate statistical model for speciated nitrogen. We will compare our multivariate model to the customary univariate model for total nitrogen, which neglects to differentiate between method of reporting TN, and highlight the differences between the two models in terms of estimating total nitrogen and inferring about important environmental drivers.

2.1. *Multivariate hierarchical data-fusion model for speciated nitrogen.* The multivariate model is specified hierarchically following the traditional form

$$[data|process, parameters][process|parameters][parameters].$$

Here, $[data|process, parameters]$ denotes the data model which will incorporate measurement error and detection limits of the measurable speciated nitrogen, given the process and parameters. The process model $[process|parameters]$ captures the true latent process of the distribution of speciated nitrogen and their dependencies. Lastly, $[parameters]$ denotes the distribution for all model parameters that specify the data model and process of speciated nitrogen. Given this hierarchical specification, model inference is commonly obtained in a Bayesian framework.

Process model. We begin by specifying the process model for speciated nitrogen given parameters. Let $\mathbf{Y}_i$ denote the vector of latent variables for lake i where $\mathbf{Y}_i = (\mathrm{DOPN}_i, \mathrm{NH}_{4,i}, \mathrm{NO_2NO}_{3,i})'$. We model these variables jointly on the log-scale to account for right skewness in the variables and to ensure nonnegative values of speciated nitrogen. In addition, the joint-model specification allows for possible dependence between these species of nitrogen at the lake level. We refer to $\mathbf{Y}_i$ as the vector of *true* values of speciated nitrogen for lake i. Given the parameters $\mathbf{B}$ and $\mathbf{\Sigma}$, the process model is specified as

$$\log(\mathbf{Y}_i) \sim \mathrm{MVN}(\mathbf{B}\mathbf{X}_i, \mathbf{\Sigma}), \tag{3}$$

where $\mathbf{X}_i$ is a length p vector of lake specific covariates, $\mathbf{B}$ is a $3 \times p$ matrix of coefficients and $\mathbf{\Sigma}$ is a 3×3 matrix capturing the dependence in the latent variables within lake and variability across lake that is not accounted for by the regression. Importantly, this process-model specification enables resolved species-specific inference, both in terms of the environmental drivers of nitrogen through the parameter matrix $\mathbf{B}$ and the unexplained variability in nitrogen across lake through the covariance matrix $\mathbf{\Sigma}$.

Data model. Next, given the process and parameters, we specify the data model. We begin with a general form for the model and then specify modifications important to our application. Let r denote replicate observation for lake i, where $r = 1, \ldots, R_i$ with R_i equalling the number of observations. Then, define $\mathbf{Z}_{i(r)}$ as the vector of measurable speciated nitrogen data, where $\mathbf{Z}_{i(r)} = (\mathrm{TN}_{i(r)}, \mathrm{TKN}_{i(r)}, \mathrm{NH}_{4,i(r)}, \mathrm{NO_2NO}_{3,i(r)})'$. Recall from (1) and (2) that TN_i and TKN_i are composed of the distinct elements of the process model, $\mathbf{Y}_i$, and that DOPN_i is never measured directly, only as a component of both TN_i and TKN_i. Let $\boldsymbol{\eta}_i = (\eta_{i1}, \eta_{i2}, \eta_{i3}, \eta_{i4})'$ be the vector of the true values of the measurable variables for lake i. That is, η_{i1}, η_{i2}, η_{i3} and η_{i4} are the true values of TN_i, TKN_i, $\mathrm{NH}_{4,i}$ and $\mathrm{NO_2NO}_{3,i}$, respectively. Then, $\boldsymbol{\eta}_i$ can be written as a linear combination of the unique elements of $\mathbf{Y}_i$ according to the matrix $\mathbf{M}$, such that

$$\begin{pmatrix} \eta_{i1} \\ \eta_{i2} \\ \eta_{i3} \\ \eta_{i4} \end{pmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \end{pmatrix},$$

$$\boldsymbol{\eta}_i = \mathbf{M}\mathbf{Y}_i.$$

Then, the data model can be written as

$$Z_{i(r)j} = \eta_{ij} + \epsilon_{i(r)j}, \tag{4}$$

where $\epsilon_{i(r)j}$ denotes the measurement error for lake i, species j and replicate r. Here, a typical measurement error model would assume conditional independence across species and

replicate given $\boldsymbol{\eta}_i$. Marginally, dependence between replicate observations of speciated nitrogen are captured through the shared η_{ij}. A common choice for the measurement error specification is an independent normal-error model $\epsilon_{i(r)j} \sim N(0, \tau_j^2)$ where the variance, τ_j^2, is species-specific. Note that, under this specification of conditional independence and within the Bayesian framework, we can easily account for missing speciated data. For example, missing values can be treated as parameters in the model and estimated within the Markov chain Monte Carlo sampling algorithm discussed in Section 2.3.

Our analysis of nitrogen concentration in lakes using LAGOS-NE requires three modifications to this general model. First, we model the observables, $Z_{i(r)j}$, on the log-scale since exploratory data analysis found measurement error to be multiplicative rather than additive. That is, the measurement error for speciated nitrogen concentrations, TN, TKN, NH_4 and NO_2NO_3, appeared to scale with observed quantities (see Appendix A). Therefore, $Z_{i(r)j} = (\log(\mathrm{TN}_{i(r)}), \log(\mathrm{TKN}_{i(r)}), \log(\mathrm{NH}_{4,i(r)}), \log(\mathrm{NO_2NO}_{3,i(r)}))'$ and $\boldsymbol{\eta}_i = \log(\mathbf{M}\mathbf{Y}_i)$.

Second, the observables, $Z_{i(r)j}$, are subject to a lower detection limit, and these need to be accounted for in the model so not to bias our estimates of TN when leveraging information across the speciated measurements. The detection limits for the observable speciated nitrogen variables vary by lake, species and replicate. Note that the variability across replicate accounts for possible changes in technologies used to collect the data at a given lake through time. Let $L_{i(r)j}$ denote the detection limit for lake i, species j and replicate r, which are fixed and known, and given on the log-scale. Nitrogen species concentrations above the detection limit can be observed with measurement error. Concentration levels below the detection limit cannot be observed and are, typically, reported as 0 or the value of the detection limit. Under the assumption of independent and normally distributed errors, we can account for these lower detection limits by rewriting (4) as

$$
\begin{aligned}
&f\big(z_{i(r)j}|\eta_{ij}, \tau_j^2, L_{i(r)j}\big) \\
&\quad = f\big(z_{i(r)j}; \eta_{ij}, \tau_j^2\big)^{I[z_{i(r)j} > L_{i(r)j}]} F\big(L_{i(r)j}; \eta_{ij}, \tau_j^2\big)^{1-I[z_{i(r)j} > L_{i(r)j}]},
\end{aligned} \tag{5}
$$

where $f(z; \eta, \tau^2)$ is the probability density function of a normal random variable with mean η and variance τ^2 and $F(L; \eta, \tau^2)$ is its cumulative distribution function. Here, $I[z_{i(r)j} > L_{i(r)j}]$ is an indicator variable equalling 1 when $z_{i(r)j}$ is above the detection limit, 0 otherwise. Importantly, this density gives probability mass to values less than $L_{i(r)j}$.

Third, we need to account for possible nonnormal measurement error distributions. Based on our exploratory data analysis using auxiliary data (see Appendix A), the t-distribution is a more appropriate specification for the measurement error model for speciated nitrogen, as it allows for heavier tails. Under the generalized three parameter location-scale family t-distribution, we can write $\epsilon_{i(r)j} \sim t(0, \tau_j^2, \nu_j)$. This assumes a location parameter equal to 0 for each species, species-specific scale and degrees of freedom parameters, τ_j^2 and ν_j, respectively. The data model in (5) is now written

$$
\begin{aligned}
f\big(z_{i(r)j}|\eta_{ij}, \tau_j^2, \nu_j, L_{i(r)j}\big) &= f\big(z_{i(r)j}; \eta_{ij}, \tau_j^2, \nu_j\big)^{I[z_{i(r)j} > L_{i(r)j}]} \\
&\quad \times F\big(L_{i(r)j}; \eta_{ij}, \tau_j^2, \nu_j\big)^{1-I[z_{i(r)j} > L_{i(r)j}]},
\end{aligned} \tag{6}
$$

where $f(z; \eta, \tau^2, \nu)$ is the probability density function of a t-distributed random variable with location parameter η, scale parameter τ^2 and degrees of freedom ν. Here, $F(L; \eta, \tau^2, \nu)$ is its cumulative distribution function.

Parameter model. To complete the hierarchical model, we specify prior distributions to each of the model parameters. The parameters include the coefficient parameters, the covariance

matrix and the measurement error variances. For example, independent normal prior distributions could be assigned to each coefficient parameter in $\mathbf{B}$, and a conjugate and noninformative inverse-Wishart distribution could be used for the covariance matrix, $\mathbf{\Sigma}$. Lastly, for a normal error distribution for $\epsilon_{i(r)j}$, a prior distribution is required for each τ_j^2. Under the t-distribution an additional prior could be assigned to each ν_j; however, this parameter is fixed in most cases. Different values of ν_j might be considered to assess the sensitivity to the prior assumption.

An important facet of this hierarchical multivariate model specification is that we are leveraging observations of multiples species across multiple reporting methods (measured vs. calculated) to inform our estimation of TN. Unless each species is observed perfectly (i.e., without any type of measurement error), observations of TN will not equate to the sum of its observed components. Additionally, each measurable species in the vector $\mathbf{Z}_{i(r)}$ is contributing to the estimate of true TN which is equal to $\sum_{l=1}^{3} Y_{il}$. Thus, we would like species that are observed with high precision to be more influential in estimating TN and those observed less precisely to be less influential. The species-specific measurement error variances, τ_j^2, control nonlinearly the relative influence of observations of species j in estimating TN. Therefore, a priori scientific information regarding measurement error for the speciated nitrogen (e.g., through the use of auxiliary data used in calibration) is valuable to aid in specifying appropriate models for $\epsilon_{i(r)j}$.

2.2. *Univariate model for total nitrogen.* We will compare the estimates from our hierarchical data-fusion model of speciated nitrogen to those from the more common univariate models for total nitrogen. Most univariate regression models for TN consider only one observation of TN per lake which is usually taken as either the median value or possibly the most recent observation. In order to objectively compare the two models, we will specify the univariate model similarly to the multivariate data fusion model presented above. That is, we will specify our univariate model hierarchically to allow for multiple replicates and to quantify the unexplained variability in TN across lake not accounted for by the regression.

Two modifications are required for the univariate model for TN. First, we will not differentiate between the two methods of reported TN and, therefore, assume we have replicate observations of TN with independent and identically distributed measurement error. Second, we will not incorporate lower detection limits into the data model since this is not possible when incorporating both methods of reporting. Thus, the detected differences between the two models can be attributed to the fusion of the speciated nitrogen data and the incorporation of lower detection limits, both of which are novel to this work.

As with the multivariate model, we will model TN using a latent variable regression model. Specifically, let $\widetilde{Z}_{i(r)}$ denote the rth observation of log(TN) for lake i where $\widetilde{\cdot}$ denotes data and parameters of the univariate model and are analogous to those in the multivariate model above. Here,

$$\widetilde{Z}_{i(r)} = \widetilde{\eta}_i + \widetilde{\epsilon}_{i(r)},$$

where $\widetilde{\eta}_i \equiv \log(\widetilde{Y}_i)$ and $\widetilde{Y}_i$ is the true latent value of TN for lake i modeled as

$$\log(\widetilde{Y}_i) \sim \mathrm{N}(\mathbf{X}_i'\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}^2).$$

Here, $\widetilde{\boldsymbol{\beta}}$ will capture the relationship between TN and the environmental covariates and $\widetilde{\sigma}^2$ accounts for the variation between lakes not accounted for by the covariates. Analogous to above, independent normal prior distributions will be assigned to $\widetilde{\boldsymbol{\beta}}$ and an inverse gamma distribution to $\widetilde{\sigma}^2$. We will employ the same measurement error distribution model for $\widetilde{\epsilon}_{i(r)}$ as in the multivariate case, with hyperparameters $\widetilde{\tau}^2$ and $\widetilde{\nu}$ under the t-distribution.

2.3. *Model inference*. Inference for the speciated-nitrogen model is obtained in a Bayesian framework. Samples from the full posterior distribution are obtained using Markov chain Monte Carlo and a hybrid Metropolis-within-Gibbs sampling algorithm. Specifically, both $\mathbf{B}$ and $\mathbf{\Sigma}$ (and $\widetilde{\boldsymbol{\beta}}$ and $\widetilde{\sigma}^2$) can be sampled directly from their full posterior distributions. Samples of the latent process, $\mathbf{Y}$, require a Metropolis–Hastings sampling algorithm. A similar Metropolis-within-Gibbs sampling algorithm was employed for fitting the univariate model for total nitrogen.

Inference includes full posterior distributions of $\mathbf{Y}_i$. That is, the distributions of the true values of DOPN, NH_4 and NO_2NO_3 for each lake as well as estimates of the covariance between these three speciated nitrogen forms. From the fitted model we can also use composition sampling to obtain full posterior distributions of TN as well as the proportion of dissolved inorganic nitrogen for each lake. Importantly, estimates of TN, computed as the sum of DOPN, NH_4 and NO_2NO_3, are ensured to adhere to the proper inequalities between the speciated nitrogen and TN. Posterior-mean estimates of these quantities as well as estimates of uncertainty can be shown for all lakes, regardless of method of reporting.

Additionally, the data-fusion model can be compared to the univariate model based on posterior mean and standard deviation estimates for TN. Differences in the parameter estimates of $\mathbf{B}$ and $\mathbf{\Sigma}$ as well as $\widetilde{\boldsymbol{\beta}}$ and $\widetilde{\sigma}^2$ can also be investigated across species and model.

3. Modeling TN across the northeast US.

3.1. *Speciated nitrogen data*. The lake data used in this analysis are available from LAGOS-NE v. 1.087.3. which contains data for over 50,000 lakes in the study regions. Nitrogen data, however, are only available for approximately 2500 lakes. To minimize possible temporal variation and seasonality in speciated nitrogen, we limited our analysis to observations obtained during the late summer months (July, August, September) between 1980 and 2012. Nitrogen is most commonly observed during summer months to avoid freezing and semifreezing conditions, and the later summer months have less annual variation in nitrogen concentration (Wetzel (2001)). For each lake, we included all observations of speciated nitrogen that were collected within five years of the most recent observation.

We limited the window of time between observations to justify the assumptions of replicate observations of speciated nitrogen, since temporal trends in nutrient concentrations as well as methodological changes in measurement over time could introduce additional uncertainty into the model (see Section 4 for further discussion).

Following the data requirements above, our analysis was conducted using speciated nitrogen data from 2305 lakes across the 17-state region of LAGOS-NE. The total number of speciated nitrogen observations across all lakes was 7496. The number of measurements obtained for each lake and species varied greatly due to sampling methods and are reported in Table 2. Of the 2305 lakes, 1602 had only one reported observation of speciated nitrogen and 338 had two. The maximum number of replicate measurements was 149. Some lakes never had direct measurements of TN and, therefore, always report *calculated TN*, whereas others never had measurements of TKN and always report *measured TN*. In addition, some replicate measurements of total nitrogen for a lake contained a combination of the two methods of reporting (e.g, some measured and some calculated). In general, the number of replicates was greater for lakes where TN was measured compared to those where TN was calculated. Of the 745 lakes having one or more observation of measured TN, 27% had five or more replicates. Conversely, only 4% of the 1560 lakes with only calculated TN had five or more replicates. For the univariate model we did not differentiate between method of reporting TN.

Detection limits vary by species and method of reporting. The median and maximum detection limit values used in our analyses are given in Table 2 for each observable species.

TABLE 2
The total number of observations of each species and the number of those below their detection limit as well as the median and maximum detection limits ($\mu g/L$) across lakes

	# Observed	# Below *DL*	Median DL	Maximum DL
TN	4936	36	84	100
TKN	2560	11	100	100
NH_4	7496	2173	10	25.8
NO_2NO_3	7496	3543	10	50

The numbers of observations below the detection limit are also given in Table 2. In the dataset, observations below the detection limit are typically reported as 0, the value of the detection limit, or something in-between (e.g., half the detection limit). TN and TKN are rarely below their detection limits, whereas 29% and 47% of the observations of NH_4 and NO_2NO_3 are below their detection limits, respectively. Recall that observations of speciated nitrogen below the detection limit are still informative when estimating TN under the model specified in Section 2.1. This information, however, was not included in the univariate model introduced in Section 2.2.

The covariates included in the model consist of known drivers of lake nutrients at the individual-lake or regional scale and were obtained from LAGOS-NE GEO v. 1.05. Region-specific variables were calculated at the hydrologic unit code-8 watershed scale and include measures of hydrology (baseflow, runoff), atmospheric N deposition and watershed land use (row crop, forest and wetland categories). Lake-specific variables include lake area, maximum depth as well as lake-to-watershed area ratio as a proxy for water residence time. We also included a variable describing connectivity class (described in detail in Fergus et al. (2017)): this characterizes whether a lake is isolated or the farthest upstream feature of a watershed (*Isolated/headwater* treated as the base case), if it has an inlet stream (*DR stream*) or if it has both lakes and streams above it in the watershed (*DR lake/stream*). Each of these covariates have been shown to be related to TN concentrations in freshwater ecosystems (e.g., Knoll et al. (2015), Read et al. (2015), Soranno et al. (2015b), Collins et al. (2017)).

3.2. *Prior distributions and model fitting.* Prior distributions were assigned to each of the model parameters. Independent, diffuse normal prior distributions with mean 0 and variance 10^4 were assigned to all coefficient parameters in $\mathbf{B}$ and $\widetilde{\boldsymbol{\beta}}$. The covariance matrix, $\boldsymbol{\Sigma}$ was assigned a noninformative inverse-Wishart distribution with degrees of freedom equal to 7 and an identity scale matrix and $\widetilde{\sigma}^2$ was assigned an inverse-Gamma prior with shape equal to 7 and scale equal to 1. These hyperprior values result in a fairly noninformative prior distribution for the residual variance of $\log(\mathbf{Y}_i)$.

Using auxiliary data collected by the USGS Standard Reference Sample Project and the North-Temperate Lakes LTER network, we conducted exploratory data analysis to inform the distributive assumptions of measurement error for speciated nitrogen. These data provided multiple reported measurements of total nitrogen as well as speciated nitrogen in which estimates of measurement error could be obtained. As a result of our exploratory analysis (details of which are given in Appendix A), we modeled measurement error using t-distributions. The auxiliary data provided estimates of the species-specific scale and degrees of parameter of the distributions which were obtained using maximum-likelihood estimation and numerical optimization. Whereas hyperprior distributions could also be assigned to the scale and degrees of freedom parameters of the measurement error variance, we opted to fix these parameters in our analyses using the empirical estimates. The scale and degrees of freedom parameter for TN, TKN, NH_4 and NO_2NO_3 are given in Table 3. A discussion of the sensitivity of the

TABLE 3
Maximum likelihood estimates of the parameters of the t-distribution for TN, TKN, NH_4 *and* NO_2NO_3 *obtained using numerical optimization*

	scale	df
TN	0.04	1.0
TKN	0.05	1.1
NH_4	0.07	0.7
NO_2NO_3	0.03	1.2

model with respect to these values is also provided in Appendix A). The same t-distribution specification for measurement error of TN was assumed in the univariate model.

Markov chain Monte Carlo was run for 200,000 iterations for both the multivariate data-fusion model and the univariate model. The first 20% of each chain were discarded as burn-in. To reduce dependence, every 20th iteration of each chain was retained for posterior inference. No issues of convergence were detected using standard diagnostics.

3.3. *Model inference*. The mean and 95% credible interval of the posterior distributions of the coefficient parameters are given in Table 4 for each species, DOPN, NH_4 and NO_2NO_3. Covariates that are significant, as deemed by their 95% credible intervals not including zero, are shown in **bold**. In addition, covariates that show opposing significant effects on the species of nitrogen are indicated with a $*$. For full comparison the last column of Table 4 gives the coefficient estimates of the univariate model for TN.

Previous regression models for total nitrogen have identified important environmental drivers (e.g., percent row crop or lake depth); however, there has been no consideration of how these drivers relate to the different components of TN. The relationships between the environmental drivers and speciated nitrogen concentration appear to vary in both magnitude and direction. Additionally, these varying relationships appear to differ across species in comparison with TN. For example, we see DOPN and NH_4 significantly decrease with an increase in runoff, as does TN in the univariate model, whereas NO_2N_3 significantly increases. These same discrepancies between speciated nitrogen and total nitrogen in the multivariate and univariate model are also seen in relation to maximum depth. NH_4 has a significant negative relationship with total atmospheric N deposition, while both DOPN and TN have a positive relationship with this variable. Collins et al. (2017) detected a negative relationship between total nitrogen and lake-watershed ratio, which our univariate model also identified. Our multivariate data fusion model discerned a significant negative relationship with lake-watershed ratio for only NO_2NO_3. We detected a negative relationship with percent forest, and this was consistent across all three species. This aligns with the results of Collins et al. (2017) who also detected a negative relationship between TN and percent forest. Lastly, we identified a positive relationship between the dissolved inorganic species, NH_4 and NO_2NO_3, and the indicator variable for lakes having both lakes and streams upstream in the watershed (DR-LakeStream) which is likely the result of pervasive fertilizer use and the flow of water with high nitrate concentration downstream.

The posterior mean and credible interval of the entries of the covariance matrix, $\mathbf{\Sigma}$, are given in Table 5. The diagonal elements show that the variability in speciated nitrogen not accounted for by the regression is relatively small for DOPN and NH_4 compared to NO_2NO_3. These estimates, however, tend to scale with speciated nitrogen concentration as NO_2NO_3 tends to be considerably larger than DOPN and NH_4. In addition, the residual covariance between DOPN and NH_4 and between NH_4 and NO_2NO_3 are significantly positive with

TABLE 4

Posterior mean and 95% *credible interval for the coefficient parameters of* $\mathbf{B}$ *and* $\widetilde{\boldsymbol{\beta}}$. *That is, the coefficients for* DOPN, NH_4 *and* NO_2NO_3 *in the multivariate model and for* TN *in the univariate model*. **bold** *signifies significant effects based on the* 95% *credible intervals, and* $*$ *identifies covariates with opposing significant effects on speciated nitrogen variables*

	Multivariate model			Univariate model
Covariate	DOPN	NH_4	NO_2NO_3	TN
Intercept	7.18 (7.08, 7.29)	3.57 (3.27, 3.86)	1.21 (0.62, 1.79)	7.14 (7.03, 7.26)
Baseflow	**−0.05 (−0.07, −0.02)**	**−0.16 (−0.22, −0.10)**	**−0.25 (−0.37, −0.13)**	**−0.06 (−0.08, −0.04)**
$*$ Runoff	**−0.28 (−0.30, −0.26)**	**−0.13 (−0.20, −0.07)**	**0.78 (0.65, 0.93)**	**−0.23 (−0.25, −0.20)**
$*$ Total N deposition	**0.05 (0.03, 0.07)**	**−0.17 (−0.23, −0.10)**	0.12 (−0.01, 0.25)	**0.05 (0.02, 0.07)**
% Row crop	**0.38 (0.25, 0.51)**	0.37 (0.00, 0.73)	**2.88 (2.20, 3.57)**	**0.68 (0.54, 0.82)**
% Forest	**−0.67 (−0.78, −0.57)**	**−0.73 (−1.02, −0.43)**	**−2.19 (−2.78, −1.61)**	**−0.73 (−0.85, −0.62)**
% Wetland	−0.06 (−0.25, 0.12)	**−0.82 (−1.35, −0.31)**	**−2.33 (−3.44, −1.19)**	−0.16 (−0.35, 0.04)
Lake area	0.00 (−0.02, 0.02)	0.02 (−0.02, 0.07)	**−0.11 (−0.21, −0.02)**	0.00 (−0.01, 0.02)
$*$ Max depth	**−0.29 (−0.32, −0.26)**	**−0.25 (−0.34, −0.17)**	**0.19 (0.03, 0.35)**	**−0.26 (−0.29, −0.23)**
Lake-watershed ratio	0.02 (0.00, 0.03)	0.00 (−0.04, 0.05)	**−0.42 (−0.51, −0.33)**	**−0.04 (−0.06, −0.02)**
DR stream	**0.07 (0.01, 0.12)**	0.16 (0.00, 0.32)	0.23 (−0.11, 0.55)	0.05 (−0.01, 0.11)
DR lake/stream	0.06 (0.00, 0.11)	**0.36 (0.19, 0.54)**	**0.89 (0.52, 1.23)**	**0.09 (0.03, 0.15)**

TABLE 5
Posterior mean and 95% credible interval for the entries of $\boldsymbol{\Sigma}$

	DOPN	NH_4	NO_2NO_3
DOPN	0.18 (0.17, 0.19)		
NH_4	0.08 (0.06, 0.11)	1.34 (1.21, 1.47)	
NO_2NO_3	0.02 (−0.03, 0.07)	1.19 (1.00, 1.36)	4.88 (4.41, 5.35)

correlation estimates 0.17 and 0.46, respectively. The posterior mean estimate of the variance parameter $\tilde{\sigma}^2$ in the univariate model for TN was 0.22 with credible interval (0.20, 0.23). This indicates that the variability in TN concentration between lakes not accounted for by the covariates is smaller than for the speciated components NH_4 and NO_2NO_3 and similar to that of DOPN.

Posterior mean and standard deviation estimates of TN for the lakes in Minnesota are shown in Figure 3. In general, TN is high in the southern part of the state and decreases as you go north. As expected, the posterior standard deviation estimates scale with TN. Maps of the posterior mean and standard deviation estimates of TN for all lakes used in the analysis are included in Figures 13 and 14 of Appendix B.

We investigated the distribution of the percentage of dissolved inorganic nitrogen, computed as $(NH_4 + NO_2NO_3)/TN$, across the region due to its important ecological impacts discussed in Section 1. Note that these estimates can only be obtained from the multivariate speciated nitrogen model. The posterior mean estimates of dissolved inorganic nitrogen for the lakes in Minnesota are shown in Figure 4 (see Figure 15 of Appendix B for all lakes). Large values of dissolved inorganic nitrogen relative to total nitrogen appear scattered throughout the state with clusters in the south-central, and north-central regions. The percent of dissolved inorganic nitrogen was small in much of the western and northeastern part of the state. The variability in percent dissolved inorganic nitrogen can be mostly attributed to the varying relationships between the species of nitrogen and the environmental factors. For example, both low baseflow values and high percentage of wetland area tend to result in a lower percentage of dissolved inorganic nitrogen. Additionally, a high percentage of dissolved inorganic nitrogen was estimated when the lake-to-watershed ratio was low. Lastly, we found that

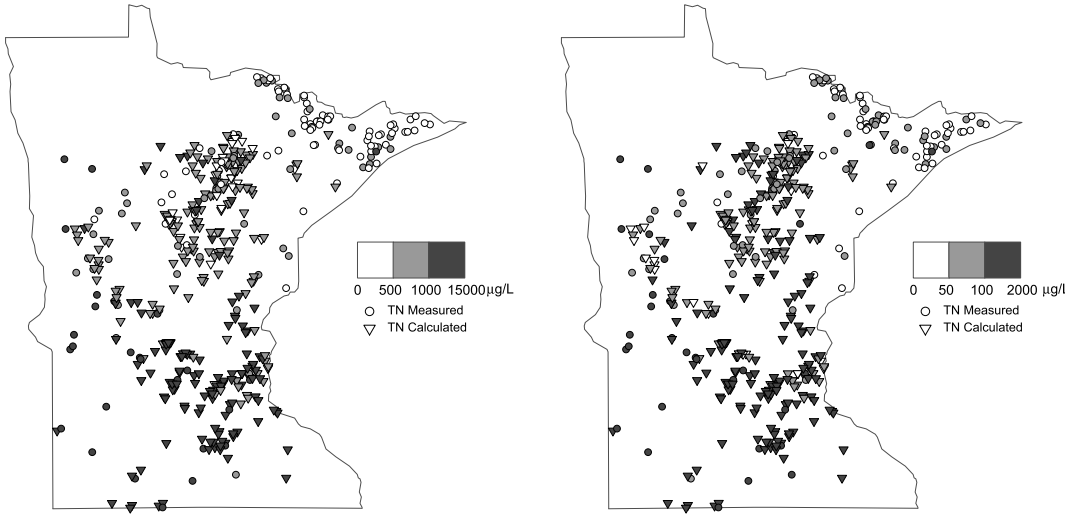

FIG. 3. *Posterior mean (left) and standard deviation (right) estimates of total nitrogen (TN) in each lake, reported in µg/L, for the state of Minnesota.*

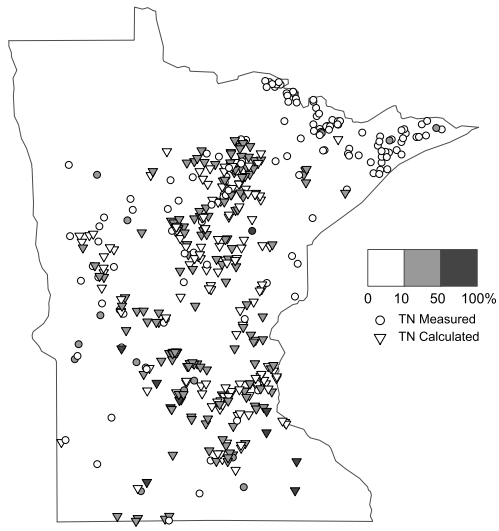

FIG. 4. *Posterior mean percentage of dissolved inorganic nitrogen, computed as* $(NH_4 + NO_2NO_3)/TN$, *in each lake for the state of Minnesota.*

the percent of dissolved inorganic nitrogen was significantly less in isolated and headwater lakes than in lakes with inlet streams or those having lakes or streams feeding into them from upstream.

3.4. *Model comparison and validation.* We compared our posterior distributions of TN from our proposed multivariate data fusion model to those of the univariate model for TN. Boxplots of the posterior estimates of the mean, standard deviation, and coefficient of variation for TN for both models and methods of reporting are shown in Figure 5. Overall, the posterior mean estimates of TN (Figure 5, left) from the univariate and multivariate models are similar, and estimates are greater for lakes where TN was calculated. This is due to the regional bias in method of reporting, where the calculated method is more prevalent in the high nutrient lakes of the upper Midwest and the measured method is more common in the low nutrient lakes of the Northeast. The posterior standard deviations of TN in Figure 5 (middle) for the two models and methods of reporting show greater uncertainty for lakes using the calculated approach which could be attributed to both uncertainty scaling with the mean and the fact that these lakes tended to have fewer number of replication observations. To address

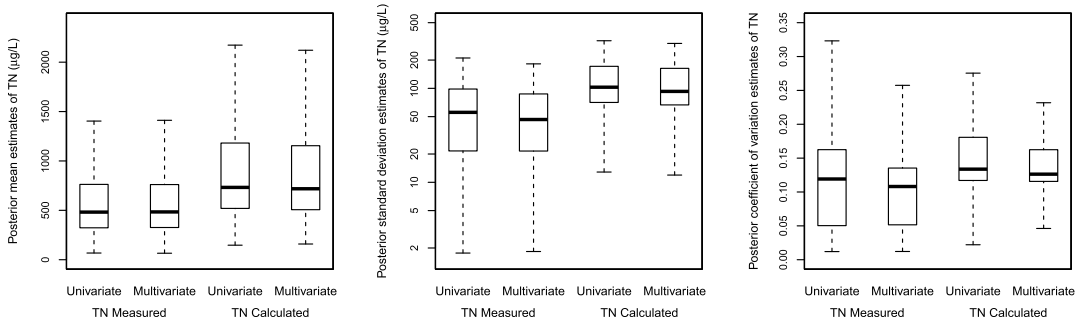

FIG. 5. *Comparison of posterior mean, standard deviation and coefficient of variation estimates of TN between the univariate and multivariate models, aggregated by method of TN reporting.* (*left*) *Boxplots of the posterior mean estimates of TN.* (*middle*) *Boxplots of the posterior standard deviations of TN.* (*right*) *Boxplots of the posterior coefficients of variation of TN.*

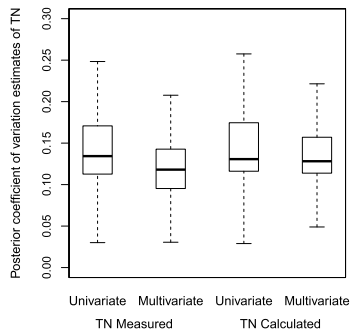

FIG. 6. *Boxplots of the posterior coefficient of variation of TN for lakes with four or fewer observations for the univariate and multivariate models, aggregated by method of TN reporting.*

these causes, we computed the posterior coefficient of variation as the ratio of the posterior standard deviation of TN relative to the posterior mean of TN for each lake (Figure 5, right). The median coefficients of variation are similar between the two methods of reporting, although measured TN has a heavier lower tail. The coefficients of variation are slightly lower for the multivariate model than the univariate model for both methods of reporting. Using only the lakes with four or fewer replicate observations, Figure 6 compares the coefficient of variation of TN across the two methods and models and found the relationships between the two methods of reporting to be very similar. Therefore, the difference between the distributions of the coefficients of variation in Figure 5 can be largely attributed to the greater number of replicate observations for lakes with TN measured. Importantly, of the lakes with fewer observations, the coefficient of variation of TN for the multivariate model was lower than the univariate model under both the measured and calculated methods. Thus, the multivariate data fusion model that incorporated observations of speciated nitrogen and accounted for the dependence between species can increase precision in the estimation of TN when fewer observations are available.

While the primary focus of this work was inferential, we used cross-validation to assess the predictive performance of our model in terms of bias and uncertainty quantification. We used 20% of the lakes as a validation set and fitted the model using the remaining 80%. The validation set contained 461 lakes and a total of 1530 speciated nitrogen observations. Due to unknown and possibly significant measurement error in reported TN, replicate observations, lower detection levels and no "true" value, we cannot assess the model using traditional metrics, such as root MSE. Instead, we relied on exploratory analysis and investigated the predictive performance graphically. Posterior predictive distributions of TN were obtained for the hold-out lakes using composition sampling. That is, we obtained posterior predictive distributions of $\sum_j Y_{ij}$ for each lake i. Figure 7 shows the posterior mean of the distribution of TN for each lake plotted against the median reported value of TN. Note that the number of observations in the validation set varied across lakes, ranging from one to 41, and some TN values from the calculated method, of which the median was taken, were computed using speciated nitrogen observations at or below their level of detection. For both methods of reporting, there is no indication of bias in these predictions of TN. Using the 95% posterior predictive intervals for TN, we also computed empirical coverage to assess our estimates of uncertainty. Importantly, these posterior predictive intervals are for the true TN for each lake, $\sum_j Y_{ij}$, and do not contain the additional uncertainty to account for error in measurement. Overall, 88% of the reported values of TN were captured by their predictive distribution. Additionally, 84% of the lakes had all of their observations of TN contained in the predictive interval, and 92% captured one or more of the reported values. Therefore, our multivariate model accurately captures uncertainty in TN.

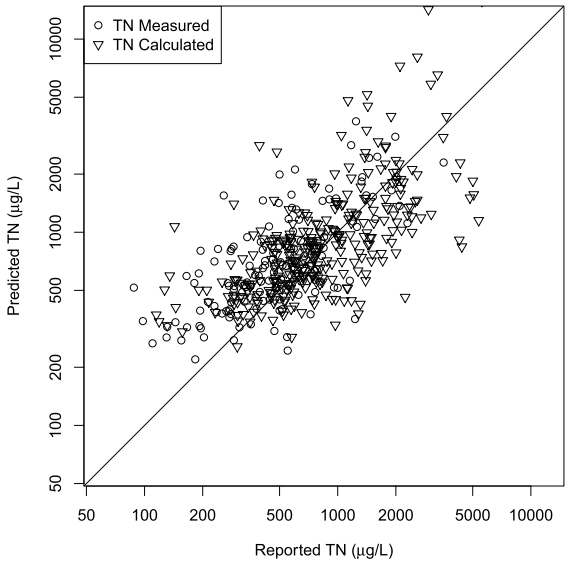

FIG. 7. *Mean of the posterior predictive distribution of TN for each lake vs. the median of the reported values, distinguished by method of reporting. Note that some calculated TN values shown were computed using speciated nitrogen observations at or below their level of detection.*

Figure 8 shows the posterior mean of speciated nitrogen, TKN, NH_4 and NO_2NO_3, for each lake plotted against the median reported value. Due to the variation in how observations below the detection limit are reported, only those lakes with a median reported value above the maximum detection limit are shown for each species. Overall, the model is able to predict TKN well out of sample. The model captures the general trend for NH_4 and NO_2NO_3, but the prediction error is much larger for these species compared to TN and TKN. This poor predictive performance is not surprising due to the large amount of variation between lakes not accounted for by the covariates for NH_4 and NO_2NO_3, as indicated by the posterior estimates of $\boldsymbol{\Sigma}$ in Table 5. In addition, information about NO_2NO_3 is only obtained from speciated observations of NO_2NO_3 and TN, whereas TN, TKN and NH_4 can leverage information from three or more speciated nitrogen observations due to their decompositions in (1) and (2). The method of reporting doesn't appear to result in any systematic over or under prediction for these species.

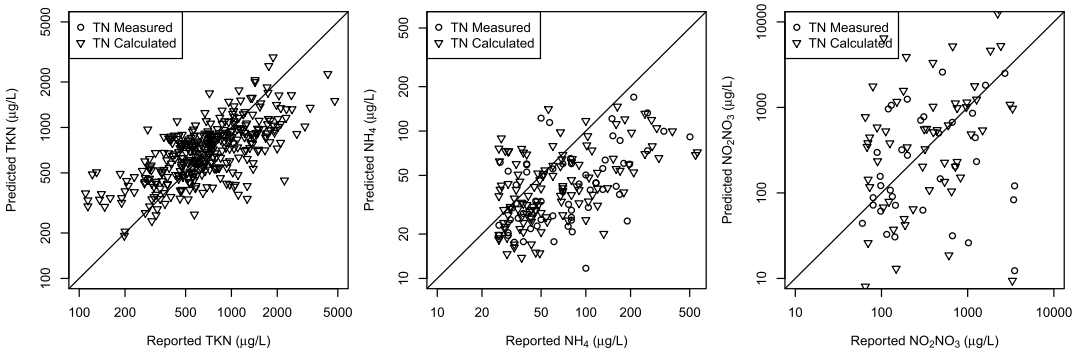

FIG. 8. *Mean of the posterior predictive distribution of speciated nitrogen for each lake vs. the median reported value, distinguished by method of reporting. For each species, only those lakes with a median reported value above their detection limit are shown.*

4. Discussion. We developed a multivariate statistical model that combined measurements of speciated nitrogen data to inform estimates of total nitrogen in lakes across continental scales. The model addressed the significant challenges posed by the data collection of lake nitrogen. Importantly, our model was able to fuse speciated nitrogen data obtained from multiple methods of reporting (direct vs. calculated) while accounting for lower detection limits that varied across species, lake and replicate observation. Due to the strong regional bias toward calculated TN in high-nutrient lakes in the upper Midwest and measured TN in low-nutrient lakes in the Northeast, failing to account for this variability would have resulted in biased estimates of TN and its environmental drivers. In addition, auxiliary data from the USGS and LTER were able to inform the measurement error distributions which were essential for leveraging multiple observations of speciated nitrogen.

The multivariate model was specified hierarchically, where the process level captured the true (latent) levels of speciated nitrogen. These latent processes, which sum to total nitrogen, enabled resolved inference with regard to identifying important environmental drivers of nitrogen. The model provided specifies-specific estimates of the relationships between nitrogen and lake- and region-specific covariates as well as captured dependence between the latent processes not accounted for by the covariates. Importantly, the variability in nitrogen concentration not accounted for by the regression was also species-specific. This species level inference is not available using the customary models for TN.

The comparison between the univariate model for TN and the proposed multivariate model indicated that the models resulted in similar estimates of TN across the region. Estimates of uncertainty from the univariate model were slightly larger than the multivariate data fusion model for both methods of reporting. Lastly, we detected a difference in the coefficient of variation estimates between lakes with TN measured vs. those with TN calculated; however, the difference could be primarily attributed to the fact that lakes where TN was calculated had fewer replicate observations.

A meaningful result of this analysis was the identification of nitrogen species-specific relationships with environmental drivers and unexplained variation, suggesting that this modeling approach will help us infer processes shaping the nitrogen cycle at broad scales and identify additional sources of variation. Even though we do not have data about transformations between species within lakes, future databases that include other important water metrics (e.g., pH) may be helpful in understanding the effects on these species. The data-fusion approach for modeling nitrogen using speciated data will become particularly critical as we extend the LAGOS-NE database to the continental US.

TN concentrations in the vast majority of lakes in the LAGOS-NE domain are temporally stationary (Oliver et al. (2017)). Further, decomposition of the components of variation in lake nutrients revealed that, while local (i.e., ecosystem-specific) temporal variation may exist for some lakes, the spatial variability is much greater than temporal variability within a macroscale extent using the LAGOS data (Soranno et al. (2019)). It is reasonable to assume that the distribution of speciated nitrogen is changing in time due to changes in environment (e.g., land use changes, changes in climate) and the interaction between lake nitrogen and other hydrologic processes. Addressing these changes in a model-based framework is challenging due to the lack of long-term data for a sufficient number of lakes (Stanley et al. (2019a)). The temporal variability in data collection alone makes this a challenging task since the LAGOS database is the compilation of extremely diverse public and private monitoring programs. Future model development includes investigating possible temporal variability within and across nitrogen species using data collected in a more limited number of lakes.

While the data fusion model developed here is very dataset and application specific, the idea of incorporating multiple sources, data types and sampling regimes easily extend to broad scientific areas. The needs of statistical models that appropriately account for the various methods of data collection and facets thereof are becoming increasingly more common as rich datasets are being compiled from a multitude of sources and programs.

APPENDIX A: DISTRIBUTIVE ASSUMPTIONS FOR MODELING MEASUREMENT ERROR

Data from the United States Geological Survey (USGS) Standard Reference Sample Project and the North-Temperate Lakes Long-Term Ecological Research network (LTER) were used to quantify the measurement error distributions for each species of nitrogen. Recall that the overall goal of this analysis was to leverage multiple observations of speciated nitrogen to improve the estimation of total nitrogen in lakes. In general, when aggregating (or averaging over) multiple measurements of total nitrogen, it is important to know the amount of uncertainty in each observation. As preliminary data analyses, we investigated the uncertainty in speciated nitrogen observations using the compilation of USGS and LTER data.

The USGS nitrogen data consist of measured values of TN, TKN and NO_2NO_3 for a collection of routine water samples as part of lab certification efforts for analyzing USGS water chemistry samples. These measurements were made by 86 water chemistry labs that participated in the project. The LTER data are derived from uniform collection and measurement methods of water from a set of 11 Wisconsin lakes over a period of 30 years. Sampling and measurement details, along with data access, are available on the North Temperate Lakes LTER web page (https://lter.limnology.wisc.edu/).

The total number of speciated nitrogen observations from the two sources were 700, 172, 245 and 1099 for TN, TKN, NH_4 and NO_2NO_3, respectively. Importantly, these samples span the range of species-specific nitrogen concentrations observed in the LAGOS data. Each water sample for which speciated nitrogen was measured also included a "most probable value," which we considered the *true* value. Using the observed and *true* values for each sample and species, we estimated the measurement error for the different species of nitrogen.

Initial investigations detected that measurement error was multiplicative rather than additive, suggesting a transformation. Therefore, the measurement error analysis was conducted on the log-transformed data. Histograms of the measurement error on the log scale are given in Figure 9 for each of the species. From our analyses we did not detect statistically significant bias for any of the species; however, the distributions appeared to be heavy-tailed. Additionally, given a water sample, our analysis of the multivariate residuals did not detect significant dependence between the different species of nitrogen.

Normal quantile-quantile plots shown in Figure 10 identify the extreme low and high values of measurement error for each species, which indicated that normally-distributed measurement error random variables (e.g., $\epsilon_{i(r)j}$ in (4)) were not appropriate for these data. From these analyses we proposed using independent t-distributions for the measurement error for each species, although other heavy-tailed distributions could be considered (e.g., Cauchy, Laplace).

Using the USGS and LTER speciated nitrogen data, we used maximum-likelihood estimation and numerical optimization methods to estimate the scale and degrees of freedom

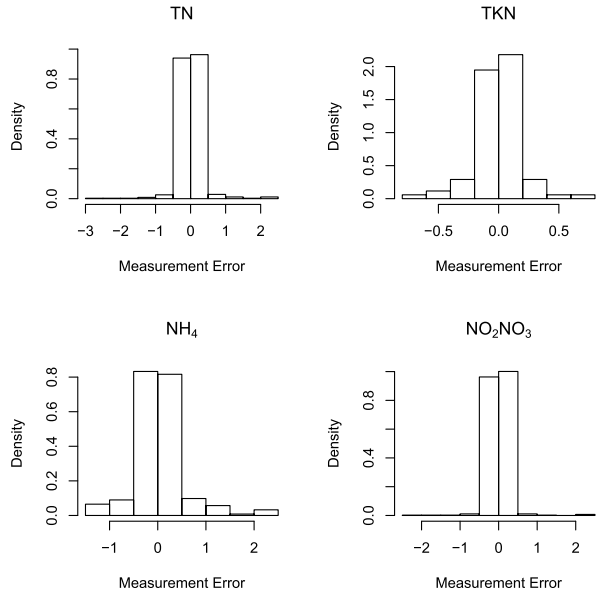

FIG. 9. *Histograms of measurement error, calculated as observed minus true values on the log-scale, for TN, TKN,* NH_4 *and* NO_2NO_3*, indicating unbiased measurement error with extremely heavy tails.*

parameters of the t-distribution for each species. Note that the location parameter was assumed to be 0 for each species. The parameter estimates for the distributions are given in Table 3 of the main text. The quantile-quantile plots using the empirical estimates of the t-distributions for each species are shown in Figure 11 and indicate that the tail behavior is better captured by the heavy-tail distributions. There was no indication of a misspecified t-distribution according to the Kolmogorov–Smirnov test for any of the speciated nitrogen variables.

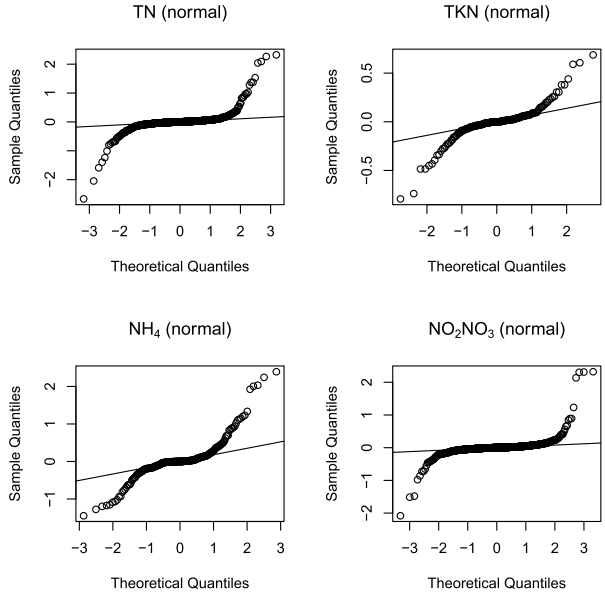

FIG. 10. *Quantile-quantile plots of the measurement error distributions of speciated nitrogen. The empirical quantiles are shown relative to quantiles of the theoretical normal distribution indicating a lack of fit.*

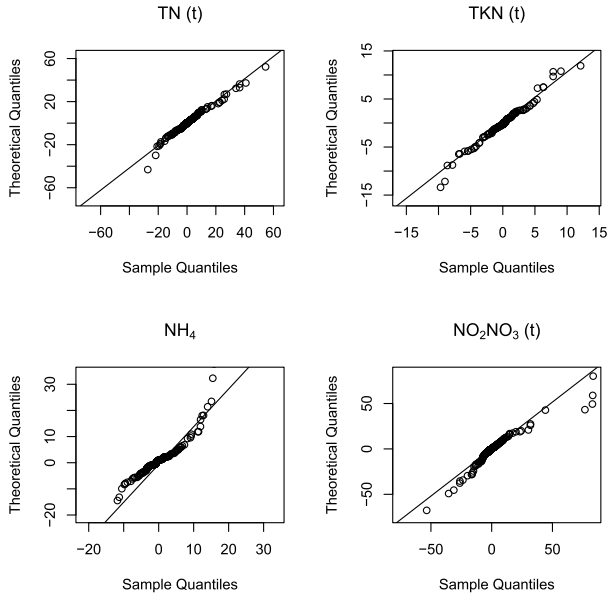

FIG. 11. *Quantile-quantile plots of the measurement error distributions of speciated nitrogen. The empirical quantiles are shown relative to quantiles of the t-distribution, parameterized using maximum likelihood estimates.*

The empirical degrees of freedom estimates for the t-distributions for each of the speciated nitrogen concentrations are close to 1, that is, the Cauchy distribution. We reconducted the Kolmogorov–Smirnov test for each species setting the degrees of freedom parameter to 1, while keeping the scale parameters set to their empirical estimates. Only the measurement error of NH_4 significantly deviated from this specification, as this species has slightly heavier tails. We assessed the sensitivity of the models as a function of these measurement error parameters by refitting the multivariate and univariate models using degrees of freedom equal to 1 for TN, TKN and NO_2NO_3 and 0.75 for NH_4. No difference was detected between the two specifications in terms of model inference or prediction. We retained the empirical estimates given in Table 3 in our analysis, as these represent our best scientific judgement regarding measurement error for speciated nitrogen.

APPENDIX B: SUPPLEMENTARY FIGURES

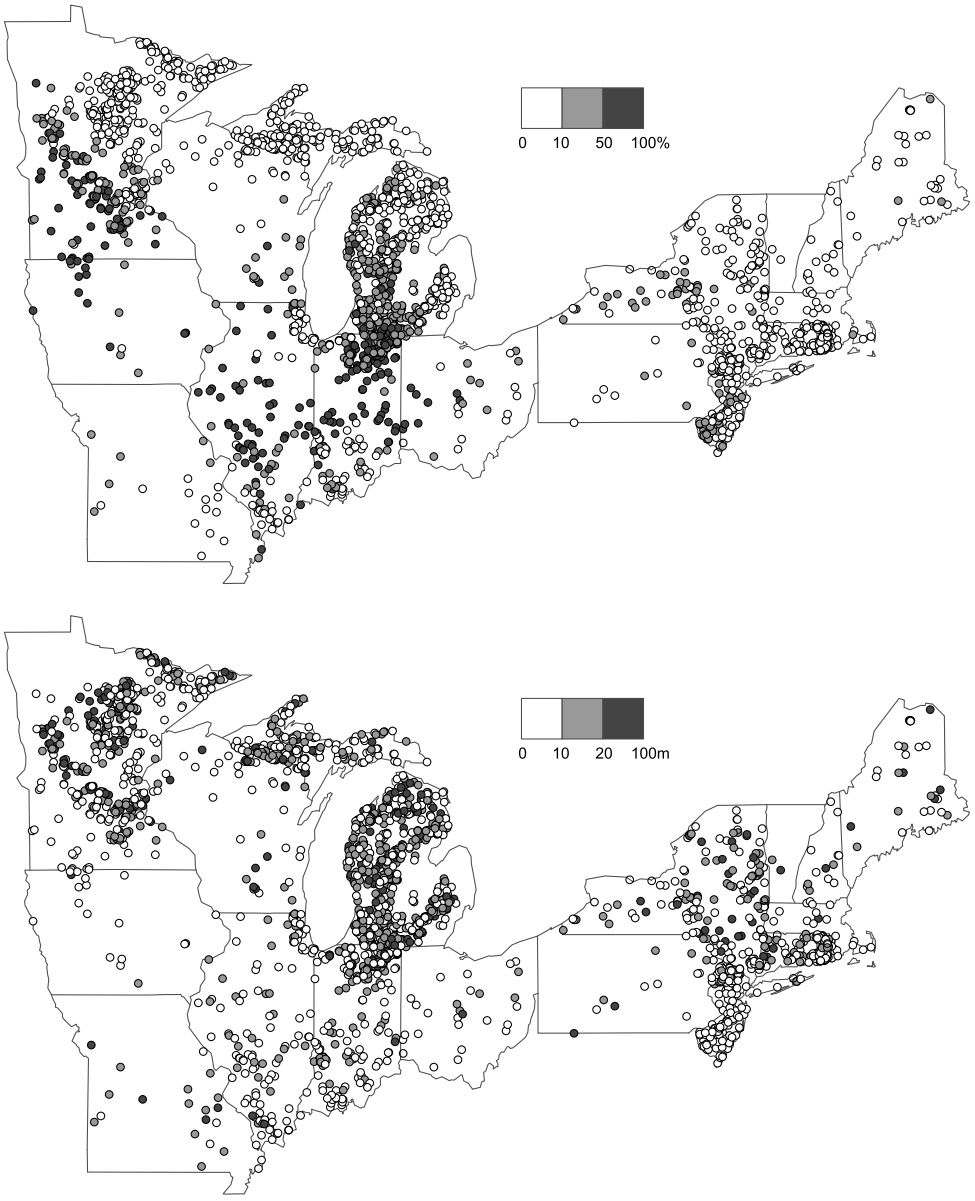

FIG. 12. (*top*) *Percent row crop agriculture in the watershed, based on the individual lake watershed* (*IWS*) *scale and* (*bottom*) *maximum lake depth, reported in m across LAGOS-NE.*

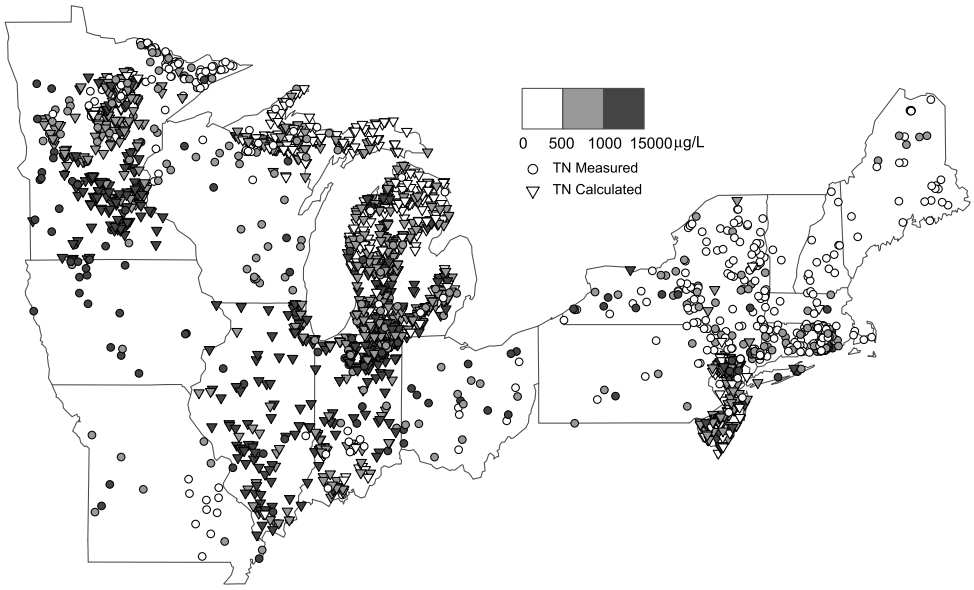

FIG. 13. *Posterior mean estimates of total nitrogen (TN), reported in µg/L.*

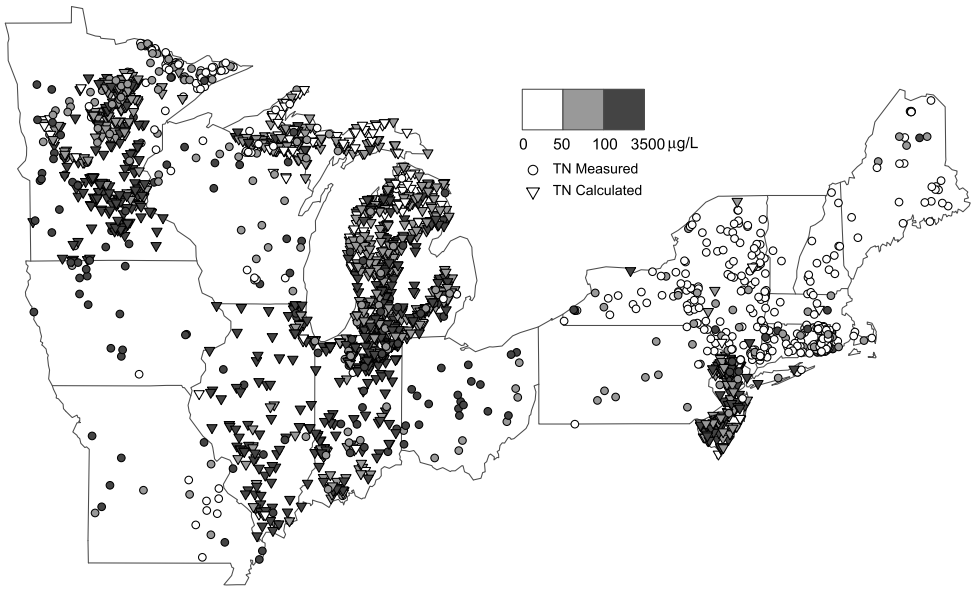

FIG. 14. *Posterior standard deviation estimates of total nitrogen (TN) in each lake.*

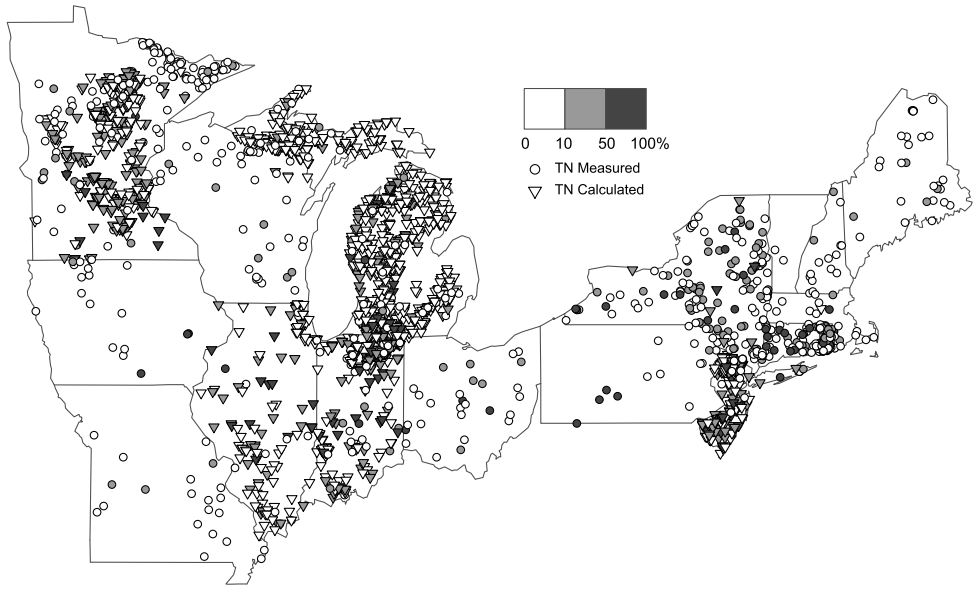

FIG. 15. *Posterior mean percentage of dissolved inorganic nitrogen, computed as* $(NH_4 + NO_2NO_3))/TN$, *in each lake.*

Acknowledgments. This research was funded by the Macrosystems Biology Program in the Emerging Frontiers Division of the Biological Sciences Directorate at the U.S. National Science Foundation (EF-1638554 and EF-1638550). Additional support was provided by NSF DEB-1440297, North Temperate Lakes LTER.

SUPPLEMENTARY MATERIAL

Supplement to "Data fusion model for speciated nitrogen to identify environmental drivers and improve estimation of nitrogen in lakes" (DOI: 10.1214/20-AOAS1371SUPP; .zip). The raw data, data formatted to fit the model, source code for the MCMC sampling algorithm, and an R script to fit the multivariate speciated nitrogen model are included as supplementary material (Schliep et al. (2020)).

REFERENCES

BERROCAL, V. J., GELFAND, A. E. and HOLLAND, D. M. (2010). A spatio-temporal downscaler for output from numerical models. *J. Agric. Biol. Environ. Stat.* **15** 176–197. MR2787270 https://doi.org/10.1007/s13253-009-0004-z

CAMERON, K., DI, H. J. and MOIR, J. (2013). Nitrogen losses from the soil/plant system: A review. *Ann. Appl. Biol.* **162** 145–173.

CHEN, M., ZENG, G., ZHANG, J., XU, P., CHEN, A. and LU, L. (2015). Global landscape of total organic carbon, nitrogen and phosphorus in lake water. *Sci. Rep.* **5** 15043. https://doi.org/10.1038/srep15043

CLARK, J. S., NEMERGUT, D., SEYEDNASROLLAH, B., TURNER, P. J. and ZHANG, S. (2017). Generalized joint attribute modeling for biodiversity analysis: Median-zero, multivariate, multifarious data. *Ecol. Monogr.* **87** 34–56.

COLLINS, S. M., OLIVER, S. K., LAPIERRE, J.-F., STANLEY, E. H., JONES, J. R., WAGNER, T. and SORANNO, P. A. (2017). Lake nutrient stoichiometry is less predictable than nutrient concentrations at regional and sub-continental scales. *Ecol. Appl.* **27** 1529–1540. https://doi.org/10.1002/eap.1545

DORIOZ, J. and FERHI, A. (1994). Non-point pollution and management of agricultural areas: Phosphorus and nitrogen transfer in an agricultural watershed. *Water Res.* **28** 395–410.

FERGUS, C. E., LAPIERRE, J., OLIVER, S. K., SKAFF, N. K., CHERUVELIL, K. S., CAREN SCOTT, K. E. W. and SORANNO, P. A. (2017). The freshwater landscape: Lake, wetland, and stream abundance and connectivity at macroscales. *Ecosphere* **8** e01911.

FUENTES, M. and RAFTERY, A. E. (2005). Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biometrics* **61** 36–45. MR2129199 https://doi.org/10.1111/j.0006-341X.2005.030821.x

GLIBERT, P. M., WILKERSON, F. P., DUGDALE, R. C., RAVEN, J. A., DUPONT, C. L., LEAVITT, P. R., PARKER, A. E., BURKHOLDER, J. M. and KANA, T. M. (2016). Pluses and minuses of ammonium and nitrate uptake and assimilation by phytoplankton and implications for productivity and community composition, with emphasis on nitrogen-enriched conditions. *Limnol. Oceanogr.* **61** 165–197.

GOBLER, C. J., BURKHOLDER, J. M., DAVIS, T. W., HARKE, M. J., JOHENGEN, T., STOW, C. A. and DE WAAL, D. B. V. (2016). The dual role of nitrogen supply in controlling the growth and toxicity of cyanobacterial blooms. *Harmful Algae* **54** 87–97. https://doi.org/10.1016/j.hal.2016.01.010

GUILLAS, S., BAO, J., CHOI, Y. and WANG, Y. (2008). Statistical correction and downscaling of chemical transport model ozone forecasts over Atlanta. *Atmos. Environ.* **42** 1338–1348.

HARPOLE, W. S., NGAI, J. T., CLELAND, E. E., SEABLOOM, E. W., BORER, E. T., BRACKEN, M. E., ELSER, J. J., GRUNER, D. S., HILLEBRAND, H. et al. (2011). Nutrient co-limitation of primary producer communities. *Ecol. Lett.* **14** 852–862.

HILKER, T., WULDER, M. A., COOPS, N. C., LINKE, J., MCDERMID, G., MASEK, J. G., GAO, F. and WHITE, J. C. (2009). A new data fusion model for high spatial-and temporal-resolution mapping of forest disturbance based on Landsat and MODIS. *Remote Sens. Environ.* **113** 1613–1627.

HILL, R. A., WEBER, M. H., DEBBOUT, R. M., LEIBOWITZ, S. G. and OLSEN, A. R. (2018). The lake-catchment (LakeCat) dataset: Characterizing landscape features for lake basins within the conterminous USA. *Freshw. Sci.* **37** 208–221. https://doi.org/10.1086/697966

KNOLL, L. B., HAGENBUCH, E. J., STEVENS, M. H., VANNI, M. J., RENWICK, W. H., DENLINGER, J. C., HALE, R. S. and GONZÁLEZ, M. J. (2015). Predicting eutrophication status in reservoirs at large spatial scales using landscape and morphometric variables. *Inland Waters* **5** 203–214.

OLIVER, S. K., COLLINS, S. M., SORANNO, P. A., WAGNER, T., STANLEY, E. H., JONES, J. R., STOW, C. A. and LOTTIG, N. R. (2017). Unexpected stasis in a changing world: Lake nutrient and chlorophyll trends since 1990. *Glob. Change Biol.* **23** 5455–5467.

PACIFICI, K., REICH, B. J., MILLER, D. A. W., GARDNER, B., STAUFFER, G., SINGH, S., MCKERROW, A. and COLLAZO, J. A. (2017). Integrating multiple data sources in species distribution modeling: A framework for data fusion. *Ecology* **98** 840–850. https://doi.org/10.1002/ecy.1710

PAERL, H. W., SCOTT, J. T., MCCARTHY, M. J., NEWELL, S. E., GARDNER, W. S., HAVENS, K. E., HOFFMAN, D. K., WILHEIM, S. W. and WURTSBAUGH, W. A. (2016). It takes two to tango: When and where dual nutrient (N & P) reductions are needed to protect lakes and downstream ecosystems. *Environ. Sci. Technol. Lett.* **50** 10805–10813.

READ, E. K., PATIL, V. P., OLIVER, S. K., HETHERINGTON, A. L., BRENTRUP, J. A., ZWART, J. A., WINTERS, K. M., CORMAN, J. R., NODINE, E. R. et al. (2015). The importance of lake-specific characteristics for water quality across the continental United States. *Ecol. Appl.* **25** 943–955.

RUNDEL, C. W., SCHLIEP, E. M., GELFAND, A. E. and HOLLAND, D. M. (2015). A data fusion approach for spatial analysis of speciated $PM_{2.5}$ across time. *Environmetrics* **26** 515–525. MR3431926 https://doi.org/10.1002/env.2369

SAUNDERS, J. F. III, YU, Y., MCCUTCHAN, J. H. JR and ROSARIO-ORTIZ, F. L. (2017). Characterizing limits of precision for dissolved organic nitrogen calculations. *Environ. Sci. Technol. Lett.* **4** 452–456.

SCHINDLER, D. (2012). The dilemma of controlling cultural eutrophication in lakes. *Proc. R. Soc. Lond., B Biol. Sci.* **279** 4322–4333.

SCHINDLER, D. W., HECKY, R. E., FINDLAY, D. L., STAINTON, M. P., PARKER, B. R., PATERSON, M. J., BEATY, K. G., LYNG, M. and KASIAN, S. E. M. (2008). Eutrophication of lakes cannot be controlled by reducing nitrogen input: Results of a 37-year whole-ecosystem experiment. *Proc. Natl. Acad. Sci. USA* **105** 11254–11258. https://doi.org/10.1073/pnas.0805108105

SCHLIEP, E. M., COLLINS, S. M., ROJAS-SALAZAR, S., LOTTIG, N. R. and STANLEY, E. H. (2020). Supplement to "Data fusion model for speciated nitrogen to identify environmental drivers and improve estimation of nitrogen in lakes." https://doi.org/10.1214/20-AOAS1371SUPP

SORANNO, P. A., BISSELL, E. G., CHERUVELIL, K. S., CHRISTEL, S. T., COLLINS, S. M., FERGUS, C. E., FILSTRUP, C. T., LAPIERRE, J.-F., LOTTIG, N. R. et al. (2015a). Building a multi-scaled geospatial temporal ecology database from disparate data sources: Fostering open science and data reuse. *GigaScience* **4** 28. https://doi.org/10.1186/s13742-015-0067-4

SORANNO, P. A., CHERUVELIL, K. S., WAGNER, T., WEBSTER, K. E. and BREMIGAN, M. T. (2015b). Effects of land use on lake nutrients: The importance of scale, hydrologic connectivity, and region. *PLoS ONE.* **10** e0135454.

SORANNO, P. A. and CHERUVELIL, K. S. (2017). LAGOS-NE-LIMNO v1.087.1: A module for LAGOS-NE, a multi-scaled geospatial and temporal database of lake ecological context and water quality for thousands of U.S. lakes: 1925-2013. Environmental Data Initiative.

SORANNO, P. A., BACON, L. C., BEAUCHENE, M., BEDNAR, K. E., BISSELL, E. G., BOUDREAU, C. K., BOYER, M. G., BREMIGAN, M. T., CARPENTER, S. R. et al. (2017). LAGOS-NE: A multi-scaled geospatial and temporal database of lake ecological context and water quality for thousands of US lakes. *GigaScience* **6** 1–22. https://doi.org/10.1093/gigascience/gix101

SORANNO, P. A., WAGNER, T., COLLINS, S. M., LAPIERRE, J.-F., LOTTIG, N. R. and OLIVER, S. K. (2019). Spatial and temporal variation of ecosystem properties at macroscales. *Ecol. Lett.* **22** 1587–1598.

STACHELEK, J. and OLIVER, S. (2017). LAGOSNE: R interface to the lake multi-scaled geospatial and temporal database. R package version 1.0.0.

STANLEY, E. H., COLLINS, S. M., LOTTIG, N. R., OLIVER, S. K., WEBSTER, K. E., CHERUVELIL, K. S. and SORANNO, P. A. (2019a). Biases in lake water quality sampling and implications for macroscale research. *Limnol. Oceanogr.* **10** e0135454.

STANLEY, E. H., SALAZAR, S. R., SCHLIEP, E. M., LOTTIG, N. R., FILSTRUP, C. T. and COLLINS, S. M. (2019b). Comparison of total nitrogen data from direct and kjeldahl-based approaches in integrated datasets. *Limnol. Oceanogr., Methods* **17** 639–649.

STOW, C. A., WEBSTER, K. E., WAGNER, T., LOTTIG, N., SORANNO, P. A. and CHA, Y. (2018). Small values in big data: The continuing need for appropriate metadata. *Ecol. Inform.* **45** 26–30.

WAGNER, T. and SCHLIEP, E. M. (2018). Combining nutrient, productivity, and landscape-based regressions improves predictions of lake nutrients and provides insight into nutrient coupling at macroscales. *Limnol. Oceanogr.* **63** 2372–2383.

WETZEL, R. G. (2001). *Limnology*: *Lake and River Ecosystems*. Gulf Professional Publishing.

WU, Z., LIU, Y., LIANG, Z., WU, S. and GUO, H. (2017). Internal cycling, not external loading, decides the nutrient limitation in eutrophic lake: A dynamic model with temporal Bayesian hierarchical inference. *Water Res.* **116** 231–240.