RESEARCH ARTICLE

Identifying and characterizing extrapolation in multivariate response data

Meridith L. Bartley[1☯*], **Ephraim M. Hanks**[1☯], **Erin M. Schliep**[2☯], **Patricia A. Soranno**[3☯], **Tyler Wagner**[4☯]

1 Department of Statistics, Pennsylvania State University, University Park, Pennsylvania, United States of America, **2** Department of Statistics, University of Missouri, Columbia, Missouri, United States of America, **3** Department of Fisheries and Wildlife, Michigan State University, East Lansing, Michigan, United States of America, **4** U.S. Geological Survey, Pennsylvania Cooperative Fish and Wildlife Research Unit, Pennsylvania State University, University Park, Pennsylvania, United States of America

☯ These authors contributed equally to this work.
* bartley@psu.edu

Abstract

Faced with limitations in data availability, funding, and time constraints, ecologists are often tasked with making predictions beyond the range of their data. In ecological studies, it is not always obvious when and where extrapolation occurs because of the multivariate nature of the data. Previous work on identifying extrapolation has focused on univariate response data, but these methods are not directly applicable to multivariate response data, which are common in ecological investigations. In this paper, we extend previous work that identified extrapolation by applying the predictive variance from the univariate setting to the multivariate case. We propose using the trace or determinant of the predictive variance matrix to obtain a scalar value measure that, when paired with a selected cutoff value, allows for delineation between prediction and extrapolation. We illustrate our approach through an analysis of jointly modeled lake nutrients and indicators of algal biomass and water clarity in over 7000 inland lakes from across the Northeast and Mid-west US. In addition, we outline novel exploratory approaches for identifying regions of covariate space where extrapolation is more likely to occur using classification and regression trees. The use of our Multivariate Predictive Variance (MVPV) measures and multiple cutoff values when exploring the validity of predictions made from multivariate statistical models can help guide ecological inferences.

OPEN ACCESS

Citation: Bartley ML, Hanks EM, Schliep EM, Soranno PA, Wagner T (2019) Identifying and characterizing extrapolation in multivariate response data. PLoS ONE 14(12): e0225715. https://doi.org/10.1371/journal.pone.0225715

Editor: Bryan C Daniels, Arizona State University & Santa Fe Institute, UNITED STATES

Received: May 7, 2019

Accepted: November 10, 2019

Published: December 5, 2019

Data Availability Statement: All data and R code used in our analysis have been compiled into a GitHub repository, https://github.com/MLBartley/MV_extrapolation. Current releases are available at https://github.com/MLBartley/MV_extrapolation/releases, and a static version of the package has been publicly archived via Zenodo (DOI: 10.5281/zenodo.3523116). All data used and created in this analysis are archived via figshare (DOI: 10.6084/m9.figshare.10093460).

Funding: Funding was provided by the US NSF Macrosystems Biology Program grants, DEB-

Introduction

The use of ecological modeling to translate observable patterns in nature into quantitative predictions is vital for scientific understanding, policy making, and ecosystem management. However, generating valid predictions requires robust information across a well-sampled system which is not always feasible given constraints in gathering and accessing data. Extrapolation is defined as a prediction from a model that is a projection, extension, or expansion of an estimated model (e.g. regression equation, or Bayesian hierarchical model) beyond the range

1638679; DEB-1638550, DEB-1638539, DEB-1638554 (EH, PS, TW and ES, https://www.nsf.gov/funding/pgm%20summ.jsp?pims%20id=503425). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

of the data set used to fit that model [1]. When we use a model fit on available data to predict a value or values at a new location, it is important to consider how dissimilar this new observation is to previously observed values. If some or many covariate values of this new point are dissimilar enough from those used when the model was fitted (i.e. either because they are outside the range of individual covariates or because they are a novel combination of covariates) predictions at this point may be unreliable. Fig 1, adapted from work by Filstrup et al. [2], illustrates this risk with a simple linear regression between the log transformed measurements of total phosphorous (TP) and chlorophyll *a* (Chl a) in U.S. lakes. The data shown in blue were used to fit a linear model with the estimated regression line shown in the same color. While the selected range of data may be reasonably approximated with a linear model, the linear trend does not extend into more extreme values, and thus our model and predictions are no longer appropriate.

While ecologists and other scientists know the risks associated with extrapolating beyond the range of their data, they are often tasked with making predictions beyond the range of the available data in efforts to understand processes at broad scales, or to make predictions about the effects of different policies or management actions in new locations. Forbes and Carlow [3] discuss the double-edged sword of supporting cost-effective progress while exhibiting caution for potential misleading results that would hinder environmental protections. They outline the need for extrapolation to balance these goals in ecological risk assessment. Other works [4–6]

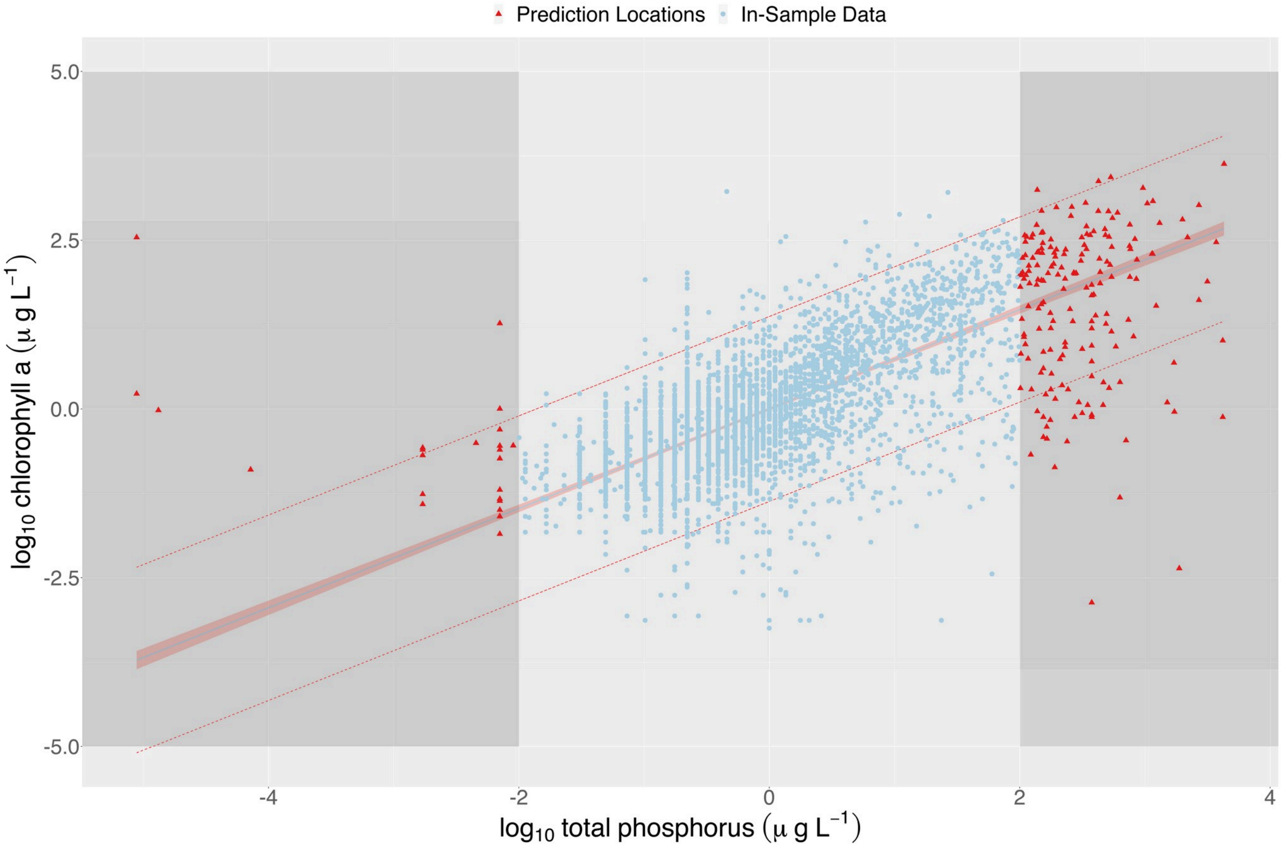

Fig 1. Fit of Chl a–TP relationship for inland lakes using linear regression. A 95% confidence interval of the mean is included around the regression line. Dashed red lines represents the 95% prediction interval. Areas shaded in darker grey indicate regions of extrapolation (using the maximum leverage value (h_{ii}) to identify the boundaries).

https://doi.org/10.1371/journal.pone.0225715.g001

explore strategies for addressing the problem of ecological extrapolation, often in space and time, across applications in management tools and estimation practices. Previous work on identifying extrapolation includes Cook's early work on detecting outliers within a simple linear regression setting [7] and recent extensions to GLMs and similar models by Conn et al. [8]. The work of Conn et al. defines extrapolation as making predictions that occur outside of a generalized independent variable hull (gIVH), defined by the estimated predictive variance of the mean at observed data points. This definition allows for predictions to be either interpolations (inside the hull) or extrapolations (outside the hull).

However, the work of Conn et al. [8] is restricted to univariate response data, which does not allow for the application of these methods to multivariate response models. This is an important limitation because many ecological and environmental research problems are inherently multivariate in nature. Elith and Leathwick [9] note the need for additional extrapolation assessments of fit in the context of using species distribution models (SDMs) for forecasting across different spatial and temporal scales. Mesgaran et al. [10] developed a new tool for identifying extrapolation using the Mahalanobis distance to detect and quantify the degree of dissimilarity for points either outside the univariate range or forming novel combinations of covariates.

In our paper, we present a general framework for quantifying and evaluating extrapolation in multivariate response models that can be applied to a broad class of problems. Our approach may be succinctly summarized as follows:

1. Fit an appropriate model to available multi-response data.
2. Choose a numeric measure associated with extrapolation that provides a scalar value in a multivariate setting.
3. Choose a cutoff or range of cutoffs for extrapolation/interpolation.
4. Given a cutoff, identify locations that are extrapolations.
5. Explore where extrapolations occur. Use this knowledge to help inform future analyses and predictions.

We draw on extensive tools for measures of leverage and influential points to inform decisions of a cutoff between extrapolation and interpolation. We illustrate our framework through an application of this approach on jointly modeled lake nutrients, productivity, and water clarity variables in over 7000 inland lakes from across the Northeast and Mid-west US.

Predicting lake nutrient and productivity variables

Inland lake ecosystems are threatened by cultural eutrophication, with excess nutrients such as nitrogen (N) and phosphorus (P) resulting in poor water quality, harmful algal blooms, and negative impacts to higher trophic levels [11]. Inland lakes are also critical components in the global carbon (C) cycle [12]. Understanding the water quality in lakes allows for informed ecosystem management and better predictions of the ecological impacts of environmental change. Water quality measurements are collected regularly by federal, state, local, and tribal governments, as well as citizen-science groups trained to sample water quality.

The LAGOS-NE database is a multi-scaled geospatial and temporal database for thousands of inland lakes in 17 of the most lake-rich states in the eastern Mid-west and the Northeast of the continental United States [13]. This database includes a variety of water quality measurements and variables that describe a lake's ecological context at multiple scales and across multiple dimensions (such as hydrology, geology, land use, and climate).

Wagner and Schliep [14] jointly modelled lake nutrient, productivity, and clarity variables and found strong evidence these nutrient-productivity variables are dependent. They also found that predictive performance was greatly enhanced by explicitly accounting for the multivariate nature of these data. Filstrup et al. [2] more closely examined the relationship between Chl a and TP and found nonlinear models fit the data better than a log-linear model. Most notably for this work, the relationship of these variables differ in the extreme values of the observed ranges; while a linear model may work for a moderate range of these data it is imperative that caution is shown before extending results to more extreme values (i.e., to extremely nutrient-poor or nutrient-rich lakes).

In this study, following Wagner and Schliep, we consider four variables: total phosphorous (TP), total nitrogen (TN), Chl a, and Secchi disk depth (Secchi) as joint response variables of interests. Each lake may have observations for all four of these variables, or only a subset. Fig 2 shows response variable availability (fully observed, partially observed, or missing) for each lake in the data set. A partially observed set of response variables for a lake indicates that at least one, but not all, of the water quality measures were sampled. We consider several covariates at the individual lake and watershed scales as explanatory variables including maximum depth (m), mean base flow (%), mean runoff (mm/yr), road density (km/ha), elevation (m), stream density (km/ha), the ratio of watershed area to lake area, and the proportion of forested

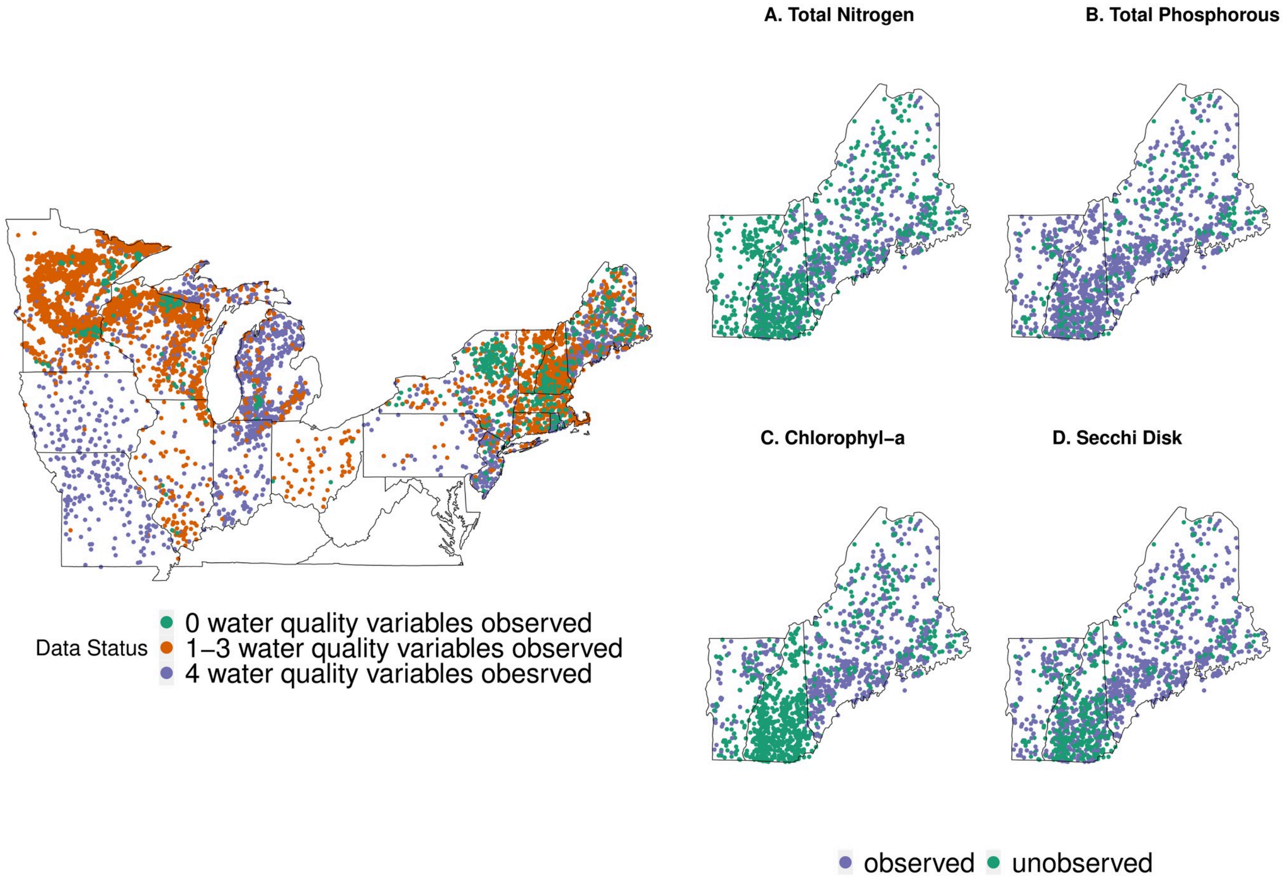
A. Total Nitrogen
B. Total Phosphorous
C. Chlorophyl–a
D. Secchi Disk
Data Status
● 0 water quality variables observed
● 1–3 water quality variables observed
● 4 water quality variables obesrved
● observed ● unobserved

Fig 2. Left: map of inland lake locations with full, partial, or missing response variables. Missing response variables are lakes where all water quality measures have not been observed, while partial status indicates only some lake response variables are unobserved. Covariates were quantified for all locations. Right: subset of data status (observed or missing) for each response variable. All spatial plots in this paper were created using the Maps package [15] in R to provide outline of US states.

https://doi.org/10.1371/journal.pone.0225715.g002

and agricultural land in each lake's watershed. One goal among many for developing this joint model is to be able to predict TN concentrations for all lakes across this region, and eventually the entire continental US. Our objective is to identify and characterize when predictions of these multivariate lake variables are extrapolations. To this end, we will review and develop methods for identifying and characterizing extrapolation in multivariate settings.

Materials and methods

Review of current work

Cook's independent variable hull. As this work builds upon the work of Cook [7] and Conn et al. [8], we start with a review of their independent variable hull (IVH) and generalized independent variable hull (gIVH) approaches. Cook's work focuses on the identification of influential points in a linear regression setting. A linear regression model is written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

where $\mathbf{y} = [y_1, \ldots, y_n]'$ denotes a vector of n univariate observed responses, $\mathbf{X}$ denotes the covariate matrix with an intercept, $\boldsymbol{\beta}$ are the covariate coeffecients, and $\boldsymbol{\epsilon}$ are independent, mean-zero normally distributed residuals. Throughout this paper, we use bold lowercase letters to denote vectors and bold uppercase letter to denote matrices. The predicted value of $\mathbf{y}$ may be calculated

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \tag{2}$$

where $\hat{\boldsymbol{\beta}}$ may be replaced with its OLS estimate ($\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$) to obtain

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{3}$$

Equivalently, the hat matrix, $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'$, when multiplied by the observed $\mathbf{y}$ vector will produce the predicted values. The predicted response for observation i can be written as a linear combination of the n response variables,

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \ldots + h_{ii}y_i + \ldots + h_{in}y_n \qquad \text{for } i = 1, \ldots, n. \tag{4}$$

The diagonal elements of the hat matrix ($h_{ii} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$) are called leverages, and while they only depend on the explanatory variables, they indicate the influence observations, y_i, have on their own predicted values, $\hat{y}_i$. A higher leverage h_{ii} indicates a higher influence of y_i in determining the model fitted response $\hat{y}_i$. This relationship means leverage values are useful quantities to explore when looking for influential points. The corresponding residual vector is $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$. Building on confidence ellipsoids for multiple coefficients, Cook's Distance, D_i, is a measure to explore the individual contribution of the i^{th} data point in a linear regression analysis. This measure may be calculated by

$$\begin{aligned} D_i &= \frac{(\hat{\boldsymbol{\beta}}_{-(i)} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}_{-(i)} - \hat{\boldsymbol{\beta}})}{ps^2} \\ &= \frac{t_i^2}{p}\left(\frac{h_{ii}}{1 - h_{ii}}\right) \end{aligned} \tag{5}$$

where p represents the number of parameters, s^2 is $\mathbf{r}'\mathbf{r}/(n - p)$, and t_i is the i^{th} studentized residual. We use $\hat{\boldsymbol{\beta}}_{-(i)}$ to indicate the estimate of the the $\boldsymbol{\beta}$ vector without the i^{th} data point. With all other values held constant, this measure increases as a function of the ratio of h_{ii} over $1 - h_{ii}$, which depends only on the design points within $\mathbf{X}$. As such, Cook defines his independent variable hull (IVH) as the smallest convex set containing all of the design points. Let h denote the maximum diagonal element of this hat matrix (i.e., $h = \max(\mathrm{diag}(\mathbf{H}))$), then a new observation, $\mathbf{x}_0$, is within this defined IVH whenever

$$\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 \le h \tag{6}$$

and predicting at a point beyond the hull will imply an extrapolation.

The hat matrix and its diagonals are useful diagnostics for finding outliers in a linear regression setting. Similarly, the Mahalanobis distance (MD) [16] can be used for identifying outliers. MD and leverage are monotonically related, as the scale-invariant squared MD may be represented by

$$\begin{aligned} MD_i^2 &= (\mathbf{x}_i - \bar{\mathbf{x}})\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})' \\ &= (l-1)\left(h_{ii} - \frac{1}{l}\right) \end{aligned} \tag{7}$$

where $\mathbf{x}_i$ is a data point (with p total covariate observations), $\bar{\mathbf{x}}$ is the mean vector for all $\mathbf{x}$ (i.e. $\bar{x}_i = \frac{1}{l}\sum_{i=1}^{l} x_i$ where l is the number of observed lakes), and $\hat{\boldsymbol{\Sigma}}$ is the sample covariance matrix. We assume $\bar{\mathbf{x}} = 0$ without loss of generality. This relationship assumes the model matrix $\mathbf{X}$ includes an intercept and makes use of the following partitioning

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} \frac{1}{l} & \mathbf{0}' \\ \mathbf{0} & \frac{1}{l-1}\hat{\boldsymbol{\Sigma}}^{-1} \end{pmatrix}. \tag{8}$$

We may work backwards from h_{ii} to obtain

$$\begin{aligned} h_{ii} &= (1; \mathbf{x}_i)\begin{pmatrix} \frac{1}{l} & \mathbf{0}' \\ \mathbf{0} & \frac{1}{l-1}\hat{\boldsymbol{\Sigma}}^{-1} \end{pmatrix}(1; \mathbf{x}_i)' \\ &= \frac{1}{l} + \frac{1}{l-1}\mathbf{x}_i\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{x}_i' \\ &= \frac{1}{l} + \frac{1}{l-1}MD_i^2. \end{aligned} \tag{9}$$

This definition remains useful without any underlying distributional assumption of the data. For example, empirically obtained quantile cutoff values can serve reasonably well as threshold for declaring outliers. However, for multivariate-normal data, the squared MD can be transformed into probabilities using a chi-squared cumulative probability distribution [17] such that points that have a very high probability of not belonging to the distribution could be classified as outliers. In either scenario, outliers can be detected using only predictor variables by calculating $\mathbf{x}_0(\mathbf{X}^{'}\mathbf{X})^{-1}\,\mathbf{x}_0$ and comparing with $\max(\mathrm{diag}(\mathbf{X}(\mathbf{X}^{'}\mathbf{X})^{-1}\,\mathbf{X}))$.

Conn's generalized IVH. The work of Cook does not immediately extend to generalized linear models (GLMs) where the assumption of Gaussian errors is relaxed. To extend to the GLM case, Conn et al. (2015) define a generalized independent variable hull (gIVH) for a generalized linear model,

$$y_i \sim f_y(g^{-1}(\mu_i)), \tag{10}$$

where f_y denotes a probability density or mass function, g gives the necessary link function, and μ_i is a linear predictor (e.g. $\mu_i = \mathbf{x}_i'\boldsymbol{\beta}$). Using Cook's IVH boundary connection to predictive variance, Conn et al. define the gIVH as the set of all predicted locations $\mathscr{L}_P$ for which

$$\mathrm{var}(\hat{y}_p|\mathbf{y}) \leq \max_{o\in\mathscr{L}_O}[\mathrm{var}(\hat{y}_o|\mathbf{y})] \tag{11}$$

where $p \in \mathscr{L}_P$, $\hat{y}_p = g^{-1}(\mathbf{x}_p\hat{\beta})$ corresponds to the mean prediction at p, $\mathscr{L}_O$ denotes the set of locations where data are observed, and $\hat{y}_o$ denotes predictions of observations at $\mathbf{x}_o \in \mathscr{L}_O$. In addition to this approach of using o and p to index observed and predicted locations, respectively, we will also in this paper use i to index the collective set of locations (i.e. $i \in \mathscr{L}_o \cup \mathscr{L}_p$). The variance of this predictive mean when a non-identity link is used may be found using the delta method which may be written as

$$\begin{aligned} \mathrm{var}(\hat{y}_i) &= \mathrm{var}(g(\hat{\mu}_i)) \\ &\approx \mathbf{\Delta}\mathrm{var}(\hat{\mu}_i)\mathbf{\Delta}', \end{aligned} \tag{12}$$

where $\mathbf{\Delta}$ is a matrix of partial derivatives of the function $g(\mu)$ with respect to its parameters, evaluated at the estimators, $\hat{\mu}$.

Prediction variance. The IVH approach of Cook's work uses only the design matrix, **X**, to calculate the hat matrix, **H**. Since the hat matrix is not always well defined for more complicated models, prediction variance may be substituted as a boundary for Conn et al.'s gIVH. This Prediction Variance (PV) approach requires the design matrix, **X**, in addition to the response variable vector, **y**. Finding the prediction variance under a univariate response model is accomplished by either direct calculation of $\mathrm{var}(\hat{\boldsymbol{y}})$ or through posterior predictive inference resulting in a single scalar value for each location.

Writing our linear predictor generally as

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} \tag{13}$$

where **X** is the design matrix and $\boldsymbol{\beta}$ is the vector of unknown parameters to be estimated, we find

$$\mathrm{var}(\hat{\boldsymbol{\mu}}) = \mathbf{X}\mathrm{var}(\hat{\boldsymbol{\beta}})\mathbf{X}'. \tag{14}$$

Under a linear model,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{15}$$

where the distribution of $\hat{\boldsymbol{\beta}}$ is

$$\hat{\boldsymbol{\beta}} \sim \mathrm{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}) \tag{16}$$

and thus $\mathrm{var}(\hat{\boldsymbol{\mu}}) = \sigma^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is proportional to the hat matrix used in Cook's IVH criteria.

While this extrapolation approach can be applied under inference in both frequentist and Bayesian approaches, we focus on a Bayesian setting in which we may calculate the prediction variance using the posterior predictive distribution of an out-of-sample observation, y_p, given the observed data, **y**. Using $[\cdot]$ to denote a probability distribution, this distribution is

$$[y_p|\mathbf{y}] = \int [y_p|\boldsymbol{\theta}][\boldsymbol{\theta}|\mathbf{y}]d\boldsymbol{\theta} \tag{17}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$ in our linear model and $[\boldsymbol{\theta}|\mathbf{y}]$ is the posterior distribution. We may approximate the posterior predictive distribution through MCMC by sampling $y_p^{(a)} \sim [y_p|\boldsymbol{\theta}^{(a)}]$ using $\boldsymbol{\theta}^{(a)}$ at each iteration ($a = 1, \ldots, A$) of the algorithm. With the posterior predictive distribution and with observed covariates, $\mathbf{x}_p$ at each new location we may calculate $\mu^{(a)} = \mathbf{x}_p'\boldsymbol{\beta}^{(a)}$ at each

MCMC iteration and Monte Carlo predictive inference can be obtained using $\hat{y}_p^{(a)} = \mu^{(a)}$ for the converged MCMC samples. The prediction variance may be approximated by

$$\hat{\text{var}}(\hat{y}_p|\mathbf{y}) = \frac{\sum_{a=1}^{A} (\hat{y}_p^{(a)} - \text{E}(\hat{y}_p|\mathbf{y}))^2}{A}. \quad (18)$$

With this sample-based calculation of prediction variance for our measure of extrapolation we can easily extend this univariate approach to the multivariate setting.

Extension to the multivariate case

Building upon this previous work, we aim to extend measures of extrapolation to handle predictions of multivariate data. We illustrate this using the inland lake nutrient and productivity data. Following the multivariate linear model developed by Wagner and Schliep [18], the joint nutrient-productivity model can be collectively written as:

$$\mathbf{y}_i = \mathbf{B}\mathbf{x}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \overset{\text{iid}}{\sim} \text{N}(\mathbf{0}, \boldsymbol{\Sigma}) \quad (19)$$

where $\mathbf{y}_i$ denotes a vector column of a matrix, $\mathbf{Y}$, where y_{ni} is the value of the n$^{\text{th}}$ lake nutrient-productivity variable for lake i. For each lake i,

$$\begin{aligned} \mathbf{y}_i &= [\text{TN}_i, \text{TP}_i, \text{CHL}_i, \text{Secchi}_i]' \\ \mathbf{y}_i &= \mathbf{B}\mathbf{x}_i + \boldsymbol{\epsilon}_i \Leftrightarrow y_{ni} = \mathbf{b}_n'\mathbf{x}_i + \epsilon_{ni} \end{aligned} \quad (20)$$

where $\mathbf{B}$ is a matrix of coefficients such that $\mathbf{b}_n'$ is a row vector where b_{nq} is the coefficient of the q^{th} predictor variable for the n^{th} lake nutrient response variable. The notation $\mathbf{x}_i$ represents a $q \times 1$ vector of predictor variables for lake i. Here, again for lake i, $\boldsymbol{\epsilon}_i \sim \text{N}(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is a n x n covariance matrix capturing the dependence between nutrient-productivity variables not accounted for by the regression. We assume multivariate errors are independent and identically distributed across lakes. Following Wagner and Schliep (2018), we take a Bayesian approach and specify priors for all model parameters.

$$\begin{aligned} b_{rc} &\overset{iid}{\sim} \text{N}(0, 100), \quad r = 1, \ldots, n, c = 1, \ldots, q \\ \boldsymbol{\Sigma} &\sim \text{IW}(\text{I}, \ \text{q} + 1) \end{aligned} \quad (21)$$

Predictions of different response types covary in multivariate models, complicating our definition of a gIVH (see Eq 11) which relies on finding a maximum univariate value. Where a univariate model would yield a scalar prediction variance (Eq 18), a multivariate model will have a prediction covariance matrix. We propose capturing the size of a covariance matrix using univariate measures. Note this is similar to A-optimality and D-optimality criteria used in experimental design [19].

Further, using our novel numeric measure of extrapolation, we aim to take advantage of the multivariate response variable information to explore when we may identify an additional observation's (i.e. covariates for a new lake location) predictions as extrapolations for all response values, jointly. We also present an approach to identify when we cannot trust a prediction for only a single response variable at either a new lake location, or a currently partially-sampled lake. The latter identification would be useful for a range of applications in ecology. For example, in the inland lakes project, one important goal is to predict TN because this essential nutrient is not well-sampled across the study extent, and yet is important for understanding nutrient dynamics and for informing eutrophication management strategies for inland lakes. In this case, to accommodate TN not being observed (i.e. sampled) as often as some other water quality variables, we can leverage the knowledge gained from samples of other water quality measures taken more often than TN (e.g. Secchi disk depth [20] is a common measure of water clarity obtained on site, while other water quality measurements require samples to be sent to a lab for analysis). We first outline our approach for identifying extrapolated new observations using a measure of predictive variance for lakes that have been fully or partially sampled and used to fit a model. Then, we describe how this approach can be applied to the prediction of TN in lakes for which it has not been sampled.

Multivariate extrapolation measures. Using available data for both complete and partial measurements of water quality at inland lake observations (here, $\mathbf{Y} = \{\mathbf{y}_o, o \in \mathscr{L}_O\}$) and corresponding covariates of these sampled locations ($\mathbf{X}$) we first fit an appropriate model to obtain estimates for parameters needed for prediction (here, $\hat{\mathbf{B}}$ and $\hat{\boldsymbol{\Sigma}}$). With these values, in addition to covariates that correspond with unsampled locations, we may either directly calculate the prediction variance or, in a Bayesian setting, simulate it via posterior predictive inference. We denote this prediction variance with $\mathbf{V}_i$ where

$$\begin{aligned} \mathbf{V}_i &= \mathrm{var}(\hat{\boldsymbol{y}}_i|\mathbf{Y}) \\ &= \mathrm{var}(\mathbf{B}\mathbf{x}_i|\mathbf{Y}). \end{aligned} \tag{22}$$

Each $\mathbf{V}_i$ is a square matrix for a sampled or unobserved location, (i.e. the combined sets of $\mathscr{L}_O$ and $\mathscr{L}_P$, respectively), with the dimensions equal to the number of response variables in the model. As in the univariate case, we propose to characterize extrapolation by comparing prediction variances of unobserved lakes with corresponding prediction variances of observed lakes. To obtain a scalar value representation of each covariance matrix we propose using the trace or determinant. In this paper, we will refer to these multivariate posterior variance (MVPV) measures for each inland lake observation with respect to how this scalar value representation is calculated:

$$\begin{aligned} \mathrm{MVPV(tr)_i} &= tr(\mathbf{V}_i) = \sum_{n=1}^{4} \mathbf{V}_i[n,n] \text{ where } \mathbf{V}_i[n,n] = n^{\text{th}} \text{ diagonal element of } \mathbf{V}_i \\ \mathrm{MVPV(D)_i} &= |\mathbf{V}_i| = \prod_{n=1}^{4} \lambda_{in} \text{ where } \lambda_{in} = n^{\text{th}} \text{ Eigenvalue of } \mathbf{V}_i \end{aligned} \tag{23}$$

The trace (tr) of an $n \times n$ square matrix $\mathbf{V}$ is defined to be the sum of the elements on the main diagonal (the diagonal from the upper left to the lower right). This does not take into account the correlation between variables and is not a scale-invariant measure. As the response variables for the inland lakes example are log transformed, we chose to explore the use of this measure for obtaining a scalar value extrapolation measure. The determinant (D) takes into account the correlations among pairs of variables and is scale-invariant. In this paper, we explore both approaches by quantifying extrapolation using our multivariate model of the LAGOS-NE lake data set by:

1. Finding the joint posterior distribution $\mathbf{B}, \Sigma|\mathbf{Y}$
2. Calculating the posterior predictive variance at in-sample lakes
3. Calculating the posterior predictive variance at out of sample lakes
4. Identifying extrapolations by comparing out of sample MVPV values to a cutoff value chosen using the in-sample values.

Conditional single variable extrapolation measures. The chosen numeric measure of MV extrapolation includes information from the entire set of responses. In the inland lake example, this could be used to identify unsampled lakes where prediction of the whole vector of response variables (TN, TP, Chl a, Secchi) are extrapolations. However, even when a joint model is appropriate, there are important scientific questions that can be answered with prediction of a single variable.

To focus on a single response variable (taken to be the n^{th} variable without loss of generality) conditioned on others, we now define the conditional multivariate predictive variance (CMVPV) as

$$\text{CMVPV}_{\text{i}} = \text{var}(\hat{y}_{ni}|\mathbf{y}_{-ni}, \mathbf{Y}) \tag{24}$$

where $\mathbf{y}_{-ni}$ are the response variables for the i^{th} lake observation being conditioned upon. With the Bayesian approach detailed above, we can get sample realizations of the conditional MVN distribution of $[y_{ni}|\mathbf{y}_{-ni}, \mathbf{Y}]$ for all MCMC iterations, however, we still need a way to obtain univariate prediction variance. While one could directly sample from $[\hat{y}_{ni}|\mathbf{y}_{-ni}, \mathbf{Y}]$, we suggest the following relationship

$$\begin{aligned} \text{var}(y_{ni}|\mathbf{Y}) &= \text{var}(\hat{y}_{ni} + \hat{\epsilon}_{ni}|\mathbf{Y}) \\ &= \text{var}(\hat{y}_{ni}|\mathbf{Y}) + \text{var}(\hat{\epsilon}_{ni}|\mathbf{Y}) \\ &= \text{var}(\hat{y}_{ni}|\mathbf{Y}) + \hat{\sigma}_n \end{aligned} \tag{25}$$

Because in our model the multivariate errors $\boldsymbol{\epsilon}_i$ are independently and identically distributed across lakes, then $\text{var}(\hat{\epsilon}_{ni}) = \sigma_n$. As σ_n is constant across all lakes, we can use either var $(y_{ni}|\mathbf{Y})$ or $\text{var}(\hat{y}_{ni}|\mathbf{Y})$ to characterize extrapolation. While the variances are different, the conclusions about extrapolation will be the same as both observed and unobserved lakes will have the same constant added.

As the inland lake data are modelled with a multivariate normal (MVN) distribution, we may use results from a conditional MVN distribution. If $\mathbf{y}_i$ is jointly normally distributed as

$$\mathbf{y}_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}) \tag{26}$$

where $\boldsymbol{\mu}_i = \mathbf{B}\mathbf{x}_i$ and if we condition the response for one nutrient measure for lake i on all other available nutrient measures for that lake then,

$$[y_{ni}|\mathbf{y}_{-ni} = \mathbf{a}, \mathbf{Y}] \sim N(\bar{\mu}, \bar{\boldsymbol{\Sigma}}) \tag{27}$$

For a lake observation that has been fully sampled for all four measures, we may compartmentalize the covariance matrix $\boldsymbol{\Sigma}$ for use in calculating the scalar values $\bar{\mu}$ and $\bar{\boldsymbol{\Sigma}}$ for $[y_{1i}|\mathbf{y}_{(2,3,4)i}]$ in the following way,

$$\boldsymbol{\Sigma} = \left[\begin{array}{c|ccc} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} & \Sigma_{14} \\ \hline \Sigma_{21} & \Sigma_{22} & \Sigma_{23} & \Sigma_{24} \\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33} & \Sigma_{34} \\ \Sigma_{41} & \Sigma_{42} & \Sigma_{43} & \Sigma_{44} \end{array}\right] \equiv \begin{bmatrix} \Sigma_{11} & \tilde{\boldsymbol{\Sigma}}_{12} \\ \tilde{\boldsymbol{\Sigma}}_{21} & \tilde{\boldsymbol{\Sigma}}_{22} \end{bmatrix} \tag{28}$$

Within this new configuration of $\boldsymbol{\Sigma}$, Σ_{11} does not change, however $\tilde{\boldsymbol{\Sigma}}_{12}$ is the submatrix of $\boldsymbol{\Sigma}$ of dimension 1× 3 containing Σ_{12}, Σ_{13}, and Σ_{14}, $\tilde{\boldsymbol{\Sigma}}_{21}$ is the submatrix of dimension 3 × 1 containing Σ_{21}, Σ_{31}, and Σ_{41}, and $\tilde{\boldsymbol{\Sigma}}_{22}$ is the submatrix of dimension 3 × 3 containing the remaining elements of $\boldsymbol{\Sigma}$. Using this partitioned $\boldsymbol{\Sigma}$ we may obtain $\bar{\mu}$ and $\bar{\Sigma}$.

$$\begin{aligned} \bar{\mu} &= \mu_{ni} - \tilde{\boldsymbol{\Sigma}}_{12}\tilde{\boldsymbol{\Sigma}}_{22}^{-1}(\mathbf{a} - \mu_{-ni}) \\ \bar{\Sigma} &= \boldsymbol{\Sigma}_{11} - \tilde{\boldsymbol{\Sigma}}_{12}\tilde{\boldsymbol{\Sigma}}_{22}^{-1}\tilde{\boldsymbol{\Sigma}}_{21} \end{aligned} \qquad (29)$$

Any of the four response variables may be considered to be variable 1 and so this general partition approach may be used for any variable conditioned on all others. The values of $\boldsymbol{\mu}_{-ni}$ and $\boldsymbol{\Sigma}$ are determined by the availability of data for the three variables we are conditioning on. These water quality measure can be fully, partially, or not observed.

In the instances where all other measures have not been observed then we may still proceed to calculate $\mathrm{var}(\hat{y}_{ni}|\mathbf{Y})$ as done for the MVPV. In order for this measure of variance to be comparable to other CMVPV values, we must add $\mathrm{var}(\hat{\epsilon}_{ni}) = \sigma_n$. This Conditional MVPV (CMVPV) measure results in a single scalar value for each location, $i \in \mathscr{L}_O \cup \mathscr{L}_P$, that may be used as outlined above to diagnose extrapolation.

Cutoffs vs continuous measures

With our selection of multivariate prediction variance measures (MVPV(tr), MVPV(D), and CMVPV) we may proceed by choosing a cutoff or range of cutoffs for delineating between extrapolation and interpretation. The role of a cutoff value or criteria in identifying and characterizing extrapolation is to delineate between prediction and extrapolation. Or rather, where (among covariate values, time, space, etc) may we expect our model to provide accurate prediction values versus where should we exhibit caution when using model-based predictions. A key decision for whether or not we label a prediction as an extrapolation (and thus identifying the location as a potentially unreliable extension of our model beyond the data) is the measure used as a boundary cutoff. Previous work [8, 21] has used the maximum prediction variance as the cutoff of the g(IVH). However, many datasets contain outliers and influential points—data locations very different from the rest of the data. Choosing a cutoff for extrapolation based on the most extreme outlier in a data set will result in a very conservative definition of extrapolation for many datasets. We thus suggest (and illustrate below) a range of extrapolation cutoffs be explored, resulting in a more complete understanding of potential extrapolation. Each cutoff value we propose is a function of the scalar valued prediction variance representations of MVPV(D or tr) and CMVPV, for observed locations ($\mathscr{L}_O$) only denoted collectively here by v_o. We examine the following cutoff options:

1. Maximum predictive variance (Cook, Conn)

$$k_{\max} = \max_{o \in \mathscr{L}_O}(v_o) \qquad (30)$$

2. Leverage-informed maximum predictive variance

$$k_{\mathrm{lev}} = \max_{o \in \mathscr{L}_{O,-\mathrm{lev}}}(v_o)$$

3. Quantile value

$$k_r = q_r(v_o, o \in \mathscr{L}_O)$$

The leverage-informed cutoff value is calculated from a set of observations in $\mathscr{L}_{O'-\mathrm{lev}'}$ where potential influential points (as determined based only on the covariate values, $\mathbf{X}$) have been removed. We suggest considering quantile-based approaches as cutoff values at the 0.99 and 0.95 quantiles of the prediction variances from observed locations. These cutoffs are less conservative than the maximum predictive variance which may also be considered the 100%

quantile value (i.e. a smaller cutoff value results in more unobserved locations identified as places where the empirical model may not be trusted).

Identifying locations as extrapolations

With the (C)MVPV values and cutoff choice in hand, determining which locations (observed/unobserved) are extrapolations is straightforward and results in a binary (yes/no) value. We refer to this delineation as our extrapolation index (e)

$$\mathrm{e}^{\mathrm{k}}{}_{p} = \begin{cases} 1 & \text{if } v_p > k \\ 0 & \text{otherwise} \end{cases} \tag{31}$$

where k represents the cutoff choice obtained using v_o. Each extrapolation index value is a function of the scalar value prediction variance representations of MVPV(D or tr) and CMVPV, for predicted locations ($\mathscr{L}_p$) only denoted collectively here by v_p. While this binary formulation allows for a simple way to determine whether or not we may diagnose a point as being an extrapolation, it does not allow for much nuance. Should a prediction with its predictive variance just beyond the boundary of the IVH be considered as untrustworthy as one with a predictive variance well beyond the boundary? We thus propose a numeric measure of extrapolation calculated by dividing predictive variance values for predicted locations by the cutoff value to generate a Relative MVPV (RMVPV) measurement:

$$\mathrm{R_k MVPV_p} = \frac{v_p}{k}. \tag{32}$$

$\mathrm{R_k}$MVPV values greater than 1 would be considered to be extrapolations, but in addition the larger the value the less trustworthy we would consider its prediction to be. This approach does not change which locations are identified as extrapolations since the binary extrapolation index as described above can be calculated from the RMVPV as

$$\mathrm{e}^{\mathrm{k}}{}_{p} = \begin{cases} 1 & \text{if } \mathrm{R_k MVPV_p} > 1 \\ 0 & \text{otherwise} \end{cases}. \tag{33}$$

Choosing IVH vs PV

With several methods of identifying extrapolations available we now provide additional guidance on choosing between various options. Cook's approach of using the maximum leverage value to define the IVH boundary may be useful for either an univariate or a joint model in a linear regression framework. However, as it depends on covariate values alone, it lacks any influence of response data. Conn et al.'s gIVH introduces the use of posterior predictive variance instead of the hat matrix to define the hull boundary in the case of a generalized model.

One possible limitation of predictive variance approaches to obtain an extrapolation index arises under certain generalized models. Models with constrained supports (i.e. binary, Poisson, etc) may exhibit decreased posterior variation when predictions are near the edges of the support. For example, in the binary case with a single covariate, if

$$\begin{aligned} y_i &\sim \mathrm{Bern}(p_i) \\ \mathrm{logit}(p_i) &= \beta_0 + \beta_1 x_i \end{aligned} \tag{34}$$

then as $x_i \rightarrow \infty$ ($x_i \leftarrow -\infty$), $p_i \rightarrow 1$ ($p_i \leftarrow 0$), and $\mathrm{var}(y_i|p_i) = p_i(1 - p_i) \rightarrow 0$. Thus, extreme points on the outside range of the observed values may have tiny predicted variance. This artificial decrease in variance may mask the identification of potentially extrapolated data points when using PV methods. Missing these extrapolations may also hinder our ability to characterize the covariate space, limiting the ability to provide reliable predictions. Thus, in models where prediction variance decreases as means go to extreme values such as Binomial, Beta, or Uniform distributions, we recommend IVH over PV approaches where this masking of extrapolation locations does not occur. We use the inland lake data set (see Predicting lake nutrient and productivity variables) to illustrate predicting joint response variables at unobserved lake locations.

Visualization and interpretation

Exploring data and taking a principled approach to identifying potential extrapolation points is often aided by visualization (and interpretation) of data and predictions. With the LAGOS data we examine spatial plots of the lakes and their locations coded by extrapolation vs prediction. Plotting this for multiple cutoff choices (as in Fig 3) is useful to explore how this choice can influence which locations are considered extrapolations. This is important from both an ecological and management perspective. For instance, if potential areas are identified as having many extrapolations this might suggest that specific lake ecosystems or landscapes have characteristics influencing processes governing nutrient dynamics in lakes not well captured by previously collected data—and thus may require further investigation.

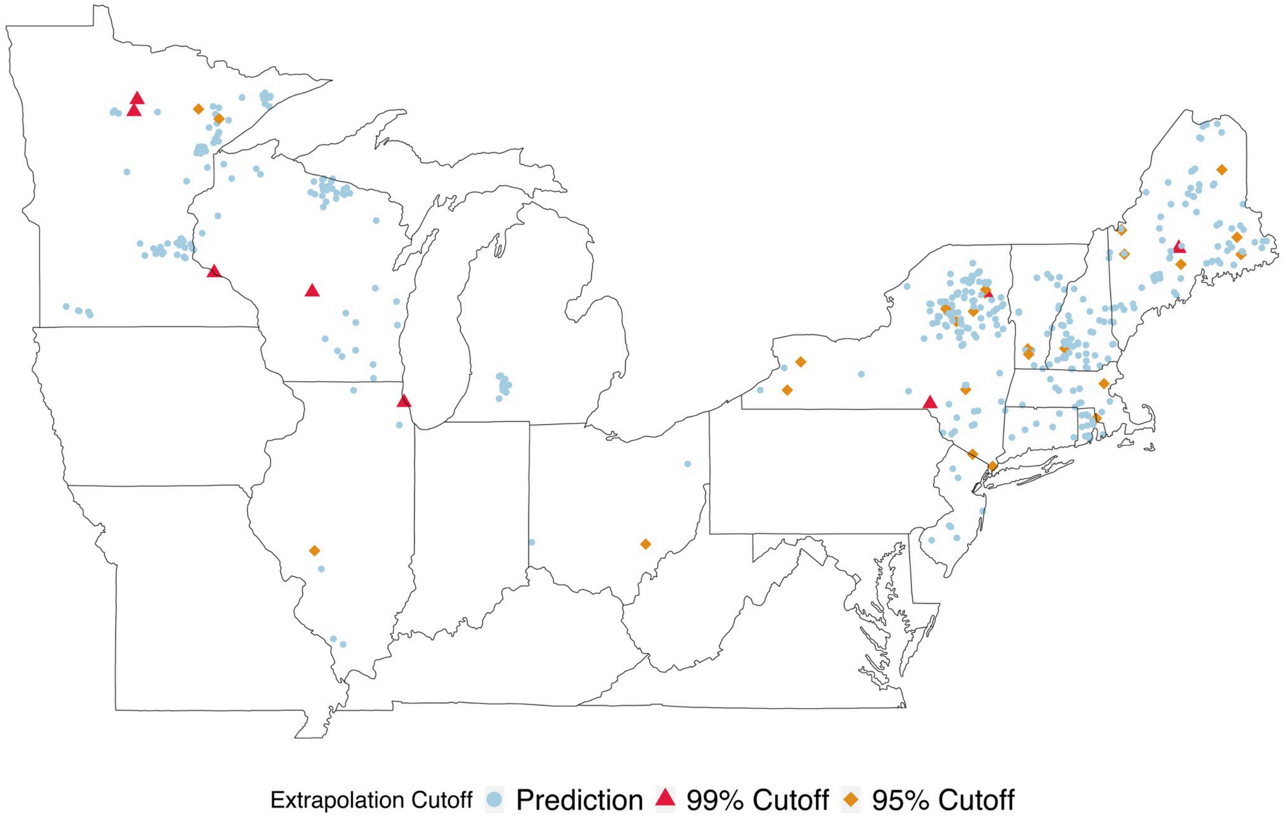

Fig 3. Identification of prediction vs extrapolation locations of LAGOS-NE lakes. Four cutoff approaches are compared and presented. Lakes in orange diamonds and red triangles indicate those where predictions were beyond the 99% and 95% cutoff values, respectively, and thus considered extrapolations. The color and shape of extrapolated lake locations are determined by which cutoff value first identifies the prediction at that location as an extrapolation.

https://doi.org/10.1371/journal.pone.0225715.g003

In addition to an exploration of possible extrapolation in physical space (through the plot in Fig 3), we also examine possible extrapolation in covariate space. Using either of the binary/numeric Extrapolation Index values, we propose a Classification and Regression Tree (CART) analysis with the extrapolation values as the response. Our classification approach allows for further insight into what covariates may be influential in determining whether a newly observed location is too dissimilar to existing ones. A CART model allows for the identification of regions in covariate space where predictions are suspect and may inform future sampling efforts as the available data has not fully characterized all lakes.

Model fitting

The joint nutrient-productivity model (see Extension to the multivariate case) was fit using MCMC in R [22]. We ran the MCMC algorithm for 20,000 iterations and used the **coda** package to analyze MCMC output and check for convergence [23]. Full conditional updates were available for all parameters (**B**, **Σ**, and **Z**) thus Gibbs updates were specified. We generated posterior predictions of lake nutrient levels across the entirety of observed and unobserved lake locations as

$$\mathbf{y}_i \sim \mathrm{N}(\mathbf{B}\mathbf{x}_i, \boldsymbol{\Sigma}) \quad i = 1, \ldots, n \tag{35}$$

and calculated multivariate prediction variance values as described in Multivariate extrapolation measures.

Results

Fitting our multivariate linear model to the 8,910 lakes resulted in most lakes' predictions remaining within the extrapolation index cutoff and thus not being identified as extrapolations. We explored the use of both trace and determinant for obtaining a scalar value representation of the multivariate posterior predictive variance in addition to four cutoff criteria. Using MVPV(tr) with these cutoffs (max value, leverage max, 0.99 quantile, and 0.95 quantile) resulted in 0, 1, 9, and 33 multivariate response predictions being identified as extrapolations, respectively. In contrast, using MVPV(D) values combined with the four cutoffs resulted in 0, 0, 8, 37 predictions identified as extrapolations. Unless all response variables are on the same scale we recommend the use of MVPV(D) over MVPV(tr). However, if a scale-invariant measure if not necessary, exploring the use of MVPV(tr) (in addition to MVPV(D) may reveal single-response variables that are of interest to researchers for further exploration using our Conditional MVPV approach. Fig 3 shows the spatial locations of lakes where the collective model predictions for TP, TN, Chl a, and Secchi depth have been identified as extrapolations using MVPV(D) combined with the cutoff measures. As the cutoff values become more conservative in nature the number of extrapolations identified increases. This figure shows the level of cutoff that first identifies a location as an extrapolation, (e.g. red squares are locations first flagged using the 99% cutoff, but they would also be included in the extrapolations found with the 95% cutoff). This increasing number of extrapolations identified highlights the importance of exploring different choices for a cutoff value. When the maximum value or the leverage-informed maximum of the predictive variance measure (k_{max} and k_{lev}) are used as cutoffs for determining when a prediction for an unsampled lake location should not be fully trusted, zero lakes are identified as extrapolations. Exploratory data analysis (see S1 Fig) indicates that for each of the lakes identified as extrapolations, the values are within the distribution of the data, with only a few exceptions. Rather than a few key variables standing out, it appears to be some combination of variables that makes a lake an extrapolation. To further characterize the type of lake more likely to be identified as an extrapolation we used a CART Model with our

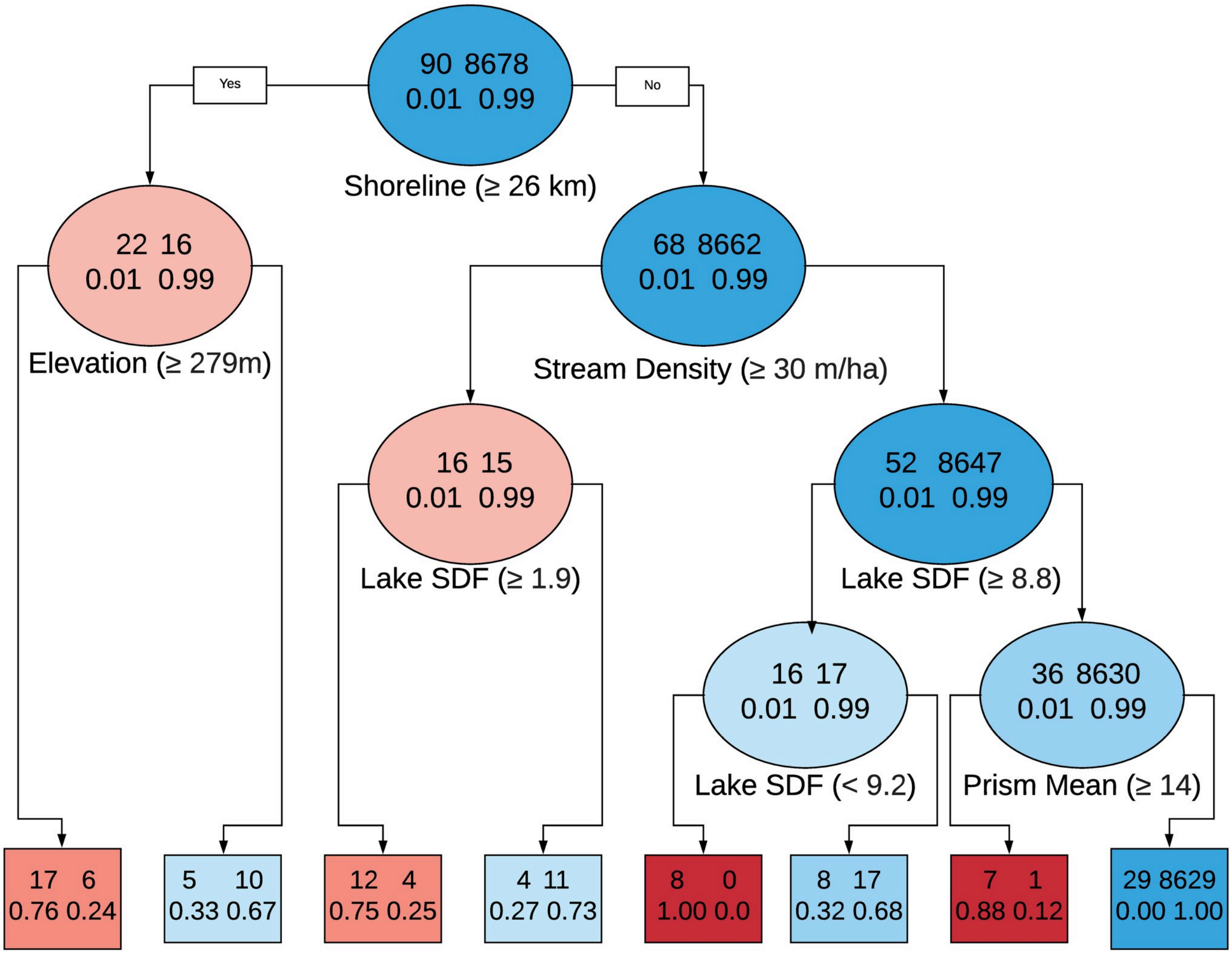

Fig 4. CART model results showing which variables may be useful in identifying extrapolations for inland lake. Each level of nodes include the thresholds and variables used to sort the data. Node color indicate whether the majority of sorted inland lake locations were identified as predictions (blue) or extrapolations (red). The first row of numbers in a node indicate the number of lakes identified as predictions (right) or extrapolations (left) that have been sorted into this node. The second row of numbers indicate the percentage of lakes that are identified as predictions (left) or extrapolation (right) with the terminal nodes (square nodes) including the percentage of records sorted by the decision tree.

https://doi.org/10.1371/journal.pone.0225715.g004

binary extrapolation index results using the MVPV(D) and the 0.95 quantile cutoff. This approach can help identify regions in the covariate space where extrapolations are more likely to occur (Fig 4). This CART analysis suggests the most important factors associated with extrapolation include shoreline length, elevation, stream density, and lake SDF. For example, a lake with a shoreline greater than 26 kilometers and above a certain elevation ($\geq$ 279 m), is likely to be identified as an extrapolation when using this model to obtain predictions. This type of information is useful for ecologists trying to model lake nutrients because it suggests lakes with these types of characteristics may behave differently than other lakes. In fact, lake perimeter, SDF, and elevation have been shown to be associated with reservoirs relative to natural lakes [24]. Although it is beyond the scope of our paper to fully explore this notion because our existing database does not differentiate between natural lakes and reservoirs, these results lend support to our approach and conclusions.

We also employed the conditional single variable extrapolation through predictive variance approach to leverage all information known about a lake when considering whether a

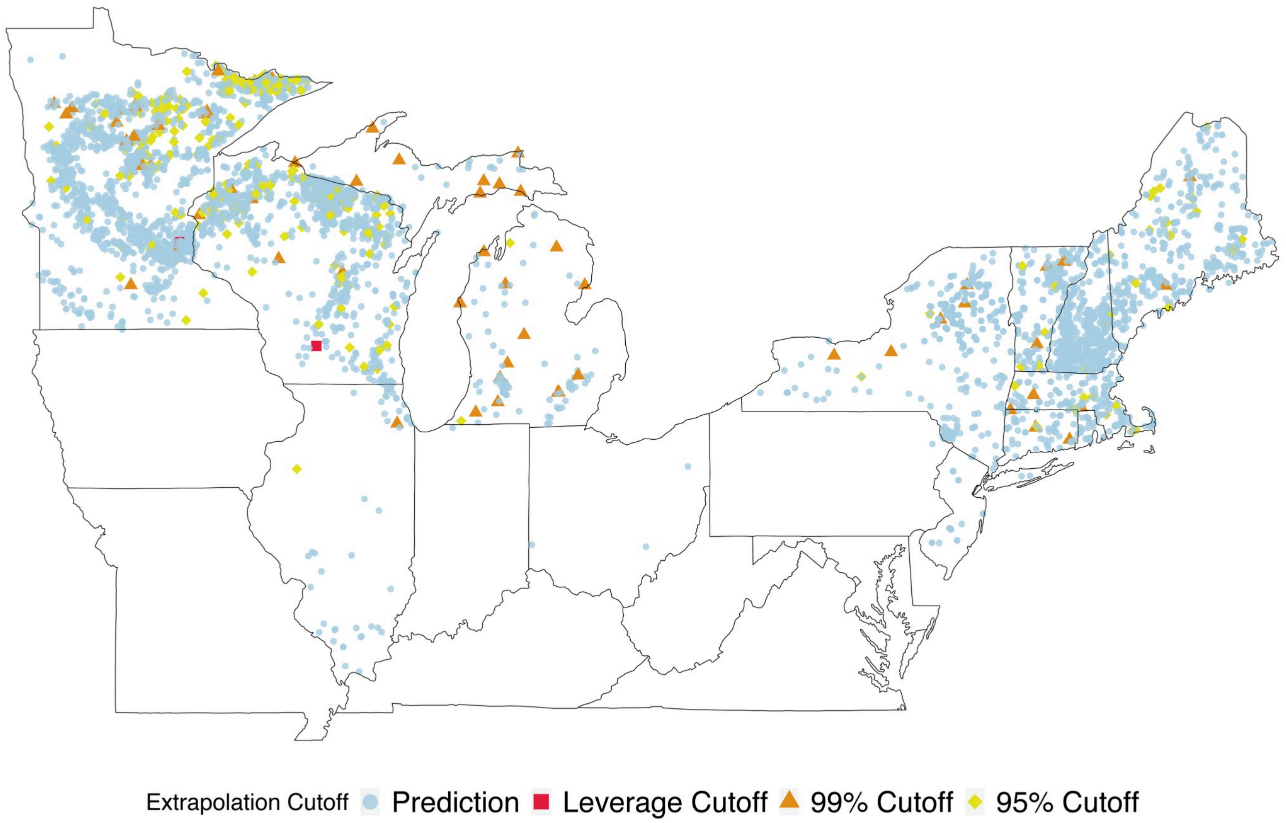

Fig 5. Identification and locations of prediction vs extrapolation of the single response variable, TN. Four cutoff approaches are compared and presented. Lakes in blue circles represent locations where TN predictions have been not been identified as extrapolations for any cutoff choice. Lakes in red squares, orange triangles, and yellow diamonds indicate those where predictions were beyond the cutoff values and thus considered extrapolations. The color and shape of extrapolated lake locations are determined by which cutoff value first identifies the prediction at that location as an extrapolation.

https://doi.org/10.1371/journal.pone.0225715.g005

prediction of a single response variable (e.g. TN, as explored here) is an extrapolation (Fig 5). These cutoffs resulted in 0, 2, 73, and 386 lake multivariate response predictions out of 5031 being identified as extrapolations. To characterize the type of lake more likely to be identified as an extrapolation we used a CART model using the 95% cutoff criterion. CART revealed the most important factors associated with extrapolation were latitude, maximum depth, and watershed to lake size ratio. Latitude may be expected as many of the lakes without measures for TN are located in the northern region. An additional visualization and table exploring extrapolation lakes and their covariate values may be found in S1 Tables.

Discussion

We have presented different approaches for identifying and characterizing potential extrapolation points within multivariate response data. Ecological research is often faced with the challenge of explaining processes at broad scales with limited data. Financial, temporal, and logistical restrictions often prevent research efforts from fully exploring an ecosystem or ecological setting. Rather, ecologists rely on predictions made on a select amount of available data that may not fully represent the breadth of a system of study. By better understanding when extrapolation is occurring scientists may avoid making unsound inferences.

In our inland lakes example we addressed the issue of large-scale predictions to fill in missing data using a joint linear model presented by Wagner and Schliep [18]. With our novel approach for identifying and characterizing extrapolation in a multivariate setting we were able to provide numeric measures associated with extrapolation (MVPV, CMVPV, R(C)MVPV) allowing for focus on predictions for all response variables or a single response variable while conditioning on others. Each of these measurements, when paired with a cutoff criterion, identify novel locations that are extrapolations. Our recommendations for visualization and interpretation of these extrapolated lakes is useful for future analyses and predictions which inform policy and management decisions. Insight into identified extrapolations and their characteristics provides additional sampling locations to consider for future work. In this analysis we found certain lakes, such as lakes located at relatively higher elevations in our study area, are more likely to be identified as an extrapolation. The available data may thus not fully represent these types of lakes, resulting in them being poorly predicted, or identified as extrapolations.

The tools outlined in this work provide novel insights into identifying and characterizing extrapolations in multivariate response settings. Further extensions of this work are available but not explored in this paper. In addition to the A- and D-optimality approaches (trace and determinant, respectively) used to obtain scalar value representations of the covariance matrices one may also explore the utility of E-optimality (maximum eigenvalue) as an additional criterion. This approach would focus on examining variance in the first principle component of the predictive variance matrix and, like the trace, this variance is not a scale-invariant measure. Our work takes advantage of posterior predictive inference under a Bayesian setting to obtain an estimate of the variance of the predictive mean response vector for each lake. However, a frequentist approach using simulation-based methods may also provide an estimate of this variance through non-parametric or parametric bootstrapping (a comparison of the two for spatial abundance estimates may be found in Hedley and Buckland [25]) and the extrapolation coefficients may be obtained through the trace and/or determinant of this variance.

This work results in identification of extrapolated lake locations as well as further understanding of the unique covariate space they occupy. The resulting caution shown when using joint nutrient models to estimate water quality variables at lakes with partially or completely unsampled measures is necessary for larger goals such as estimating the overall combined levels of varying water qualities in all US inland lakes. In addition, under- or overestimating concentrations of key nutrients such as TN and TP can potentially lead to misinformed management strategies which may have deleterious effects on water quality and the lake ecosystem. The identification of lake and landscape characteristics associated with extrapolation locations can further understanding between natural/anthropogenic sources of nutrients in lakes not well represented in the sampled population. In our database, TP is sampled more than TN, which is likely due to the conventional wisdom that inland waters are P limited, where P contributes the most to eutrophication [26]. However, nitrogen has been shown to be an important nutrient in eutrophication in some lakes and some regions [27], and may be as important to sample to fully understand lake eutrophication. Our results show it is possible to predict TN if other water quality variables are available, but it would be better if it was sampled more often.

The joint model used in this work can be improved upon in several regards; no spatial component is included, response variables are averages over several years worth of data and thus temporal variation is not considered, and data from different years are given equal weight. The model we use to fit these data may be considered to be a simple one, but the novel approach presented here may be applied to more complicated models. In a sample based approach using a Bayesian framework the MVPV and CMVPV values obtained come from the MCMC samples and are thus independent from model design choices.

Deeper understanding of where extrapolation is occurring will allow researchers to propagate this uncertainty forward. Follow up analyses using model-based predictions need to acknowledge that some predictions are less trustworthy than others. This approach and our analysis here shows that while a model may be able to produce an estimate and a confidence or prediction interval, that does not mean the truth is captured nor does the assumed relationship persist, especially outside the range of observed data. The methods outlined here will serve to guide future scientific inquiries involving joint distribution models.

Supporting information

S1 Fig. Violin plots of covariate densities and extrapolation points plotted.
(PDF)

S1 Tables. Tables of covariate values for lakes identified as extrapolations using MVPV(D) and CMVPV for TN.
(TEX)

Acknowledgments

We thank the LAGOS Continental Limnology Research Team for helpful discussions throughout the process of this manuscript. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Disclaimer

This draft manuscript is distributed solely for purposes of scientific peer review. Its content is deliberative and predecisional, so it must not be disclosed or released by reviewers. Because the manuscript has not yet been approved for publication by the US Geological Survey (USGS), it does not represent any official finding or policy.

Author Contributions

Conceptualization: Meridith L. Bartley, Ephraim M. Hanks, Erin M. Schliep, Tyler Wagner.

Data curation: Patricia A. Soranno.

Formal analysis: Meridith L. Bartley.

Funding acquisition: Patricia A. Soranno.

Methodology: Meridith L. Bartley.

Project administration: Patricia A. Soranno.

Supervision: Ephraim M. Hanks.

Visualization: Meridith L. Bartley.

Writing – original draft: Meridith L. Bartley.

Writing – review & editing: Ephraim M. Hanks, Erin M. Schliep, Patricia A. Soranno, Tyler Wagner.

References

1. Miller JR, Turner MG, Smithwick EaH, Dent CL, Stanley EH. Spatial Extrapolation: The Science of Predicting Ecological Patterns and Processes. BioScience. 2004; 54(4):310. https://doi.org/10.1641/0006-3568(2004)054%5B0310:SETSOP%5D2.0.CO;2

2. Filstrup CT, Wagner T, Soranno PA, Stanley EH, Stow CA, Webster KE, et al. Regional variability among nonlinear chlorophyll-phosphorus relationships in lakes. Limnology and Oceanography. 2014; 59(5):1691–1703.

3. Forbes V, Calow P. Extrapolation in ecological risk assessment: balancing pragmatism and precaution in chemical controls legislation. BioScience. 2002; 225(3):152–161.

4. Freckleton RP. The problems of prediction and scale in applied ecology: The example of fire as a management tool. Journal of Applied Ecology. 2004; 41(4):599–603.

5. Colwell RK, Coddington JA. Estimating Terrestrial Biodiversity through Extrapolation. Philosophical Transactions of the Royal Society B: Biological Sciences. 1994; 345(1311):101–118. https://doi.org/10.1098/rstb.1994.0091

6. Peters DPC, Herrick JE, Urban DL, Gardner RH, Breshears DD. Strategies for ecological extrapolation. Oikos. 2004; 106(3):627–636.

7. Cook RD. Detection of Influential Observation in Linear Regression. Technometrics. 1977.

8. Conn PB, Johnson DS, Boveng PL. On extrapolating past the range of observed data when making statistical predictions in ecology. PLoS ONE. 2015.

9. Elith J, Leathwick JR. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annual Review of Ecology, Evolution, and Systematics. 2009; 40(1):677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

10. Mesgaran MB, Cousens RD, Webber BL. Here be dragons: A tool for quantifying novelty due to covariate range and correlation change when projecting species distribution models. Diversity and Distributions. 2014; 20(10):1147–1159.

11. Carpenter SR, Caraco NF, Correll DL, Howarth RW, Sharpley AN, Smith VH. Carpenter_et_al-1998-Ecological_Applications. Ecological Applications. 1998; 8(January 1998):559–568.

12. Tranvik LJ, Cole JJ, Prairie YT. The study of carbon in inland waters-from isolated ecosystems to players in the global carbon cycle. Limnology and Oceanography Letters. 2018; 3(3):41–48.

13. Soranno PA, Bacon LC, Beauchene M, Bednar KE, Bissell EG, Boudreau CK, et al. LAGOS-NE: A multi-scaled geospatial and temporal database of lake ecological context and water quality for thousands of U.S. lakes. GigaScience. 2017;(March):1–22. https://doi.org/10.1093/gigascience/gix101 PMID: 29053868

14. Wagner T, Schliep EM. Combining nutrient, productivity, and landscape-based regressions improves predictions of lake nutrients and provides insight into nutrient coupling at macroscales. Limnology and Oceanography. 2018; 63(6):2372–2383.

15. Becker RA, Wilks AR, Brownrigg R, Minka TP. maps: Draw Geographical Maps. R package version 23-6. 2013;.

16. Mahalanobis PC. On the generalized distance in statistics. National Institute of Science of India. 1936;.

17. Etherington TR. Mahalanobis distances and ecological niche modelling: correcting a chi-squared probability error. PeerJ. 2019.

18. Wagner T, Schliep EM. Combining nutrient, productivity, and landscape-based regressions improves predictions of lake nutrients and provides insight into nutrient coupling at macroscales. Limnology and Oceanography. 2018.

19. Gentle JE. Matrix Algebra: Theory, Computations, and Applications in Statistics; 2007.

20. Lottig NR, Wagner T, Henry EN, Cheruvelil KS, Webster KE, Downing JA, et al. Long-term citizen-collected data reveal geographical patterns and temporal trends in lake water clarity. PLoS ONE. 2014.

21. Cook RD. Influential observations in linear regression. Journal of the American Statistical Association. 1979.

22. R Core Team. R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria URL http://wwwR-projectorg/. 2016;.

23. Vines MP, Best N, Cowles K, Karen. CODA: Convergence Diagnosis and Output Analysis for MCMC. R news. 2006; 6(1):7–11.

24. Doubek JP, Carey CC. Catchment, morphometric, and water quality characteristics differ between reservoirs and naturally formed lakes on a latitudinal gradient in the conterminous United States. Inland Waters. 2017.

25. Hedley SL, Buckland ST. Spatial models for line transect sampling. Journal of Agricultural, Biological, and Environmental Statistics. 2004.

26. Conley DJ, Paerl HW, Howarth RW, Boesch DF, Seitzinger SP, Havens KE, et al. Ecology—Controlling eutrophication: Nitrogen and phosphorus; 2009.

27. Paerl HW, Xu H, McCarthy MJ, Zhu G, Qin B, Li Y, et al. Controlling harmful cyanobacterial blooms in a hyper-eutrophic lake (Lake Taihu, China): The need for a dual nutrient (N & P) management strategy. Water Research. 2011.