Water Practice & Technology

IWA PUBLISHING

Water Practice & Technology Vol 19 No 4, 1148 doi: 10.2166/wpt.2024.068

A manifold learning perspective on surrogate modeling of nitrate concentration in the Kansas River

Nicholas Tufillaro

Gybe, PO Box 1028, Corvallis, Oregon 97333, USA
E-mail: nick@gybe.eco

NT, 0009-0006-2896-8832

ABSTRACT

A non-linear surrogate model of nitrate concentration in the Kansas River (USA) is described. The model is an (almost) Piecewise Linear response surface that provides a mean field approximation to the dynamics of the measured data for nitrate plus nitrite (target product) correlations to turbidity and chlorophyll-a concentrations (input variables). The method extends the United States Geological Survey's linear procedures for surrogate data modeling allowing for better approximations for river systems exhibiting algal blooms due to nutrient-rich source waters. The model and visualization procedures illustrated in the Kansas River example should be generally applicable to many medium-size rivers in agricultural regions.

Key words: algal bloom, Kansas River, nitrate, non-linear, surrogate modeling, water quality

HIGHLIGHT

- Non-linear surrogate modeling of nutrients is illustrate in the Kansas River that uses commonly available sensors, and initial nitrate plus nitrate calibration data, to enable future estimations of nutrients using only the turbidity and chlorophyll-a sensors.

1. INTRODUCTION

Surrogate data models have been successfully employed to infer nutrients such as nitrate and phosphorus from high-frequency (HF) time series measurements such as turbidity, discharge, and chlorophyll concentrations (Rasmussen *et al.* 2005; Williams 2021). For instance, in the Kansas River, nitrate concentrations, near the intake of the source waters for the city of Lawrence, KS, have been estimated using a log-transformed linear model with input variables of HF time series consisting of chlorophyll, turbidity, and seasonality (Foster & Graham 2016). Linear models provide useful correlations for predicting target water quality quantities, but it is also appreciated that many of the correlations between input and output variables are best described by non-linear functions (Yang & Moyer 2020). The expectation, therefore, is that non-linear modeling in surrogate data applications will more accurately track the target variables. Non-linear modeling, though, can introduce thorny issues including concerns about introducing model bias by assuming a (global) non-linear model structure that, may or may not, be inherent in the data. Additionally, topics such as over-fitting (typically addressed by 'regularization'), and sensible model extrapolations need to be considered.

All three of these topics – bias, regularization, and extrapolation – are often addressed in a non-linear setting by using the data itself to best inform a solution. One route in this direction is Bayesian inference methods and black box models (e.g., neural nets) that make weak prior assumptions. An alternative approach is the use of a domain expert to produce a parametrized model of the data with strong priors. In operational applications, black box models can be less than ideal because of their lack of transparency to the underlying biogeochemical processes. Expert models are more compelling in that regard but introduce strong priors that might not accurately describe the relevant correlations. Nevertheless, when models are connected to the physical process this inspires more confidence in practical applications, such as monitoring water quality.

In this paper, we address the points mentioned above in the context of a specific water quality surrogate model application. We examine the HF time series from the Kansas River at the USGS De Soto, Kansas, gauging station which has a long-term record of more common water quality variables, such as turbidity and chlorophyll-a concentrations, and a less common HF monitoring record for nitrate. The site is one of many where the USGS engages in estimating water quality variables following a detailed and documented surrogate modeling procedure developed and validated by the USGS over two decades (Rasmussen *et al.* 2009). Here, we illustrate how to augmente the linear surrogate modeling procedure with the introduction of an (almost) Piece-wise Linear (aPL) model that captures the global non-linearity of the data.

This surrogate modeling example sits midway between the black box and expert modeling approaches. It best falls under the rubric of manifold learning (Van Der Maaten *et al.* 2009). When embedded in a multivariate space feature space the data reveals a shape, or geometric structure, that is connected to physical processes. A backbone for this geometrical arrangement in the De Soto example can be captured with a smooth surface in two dimensions (or more generally a manifold in multiple dimensions) and serves as a useful starting point for modeling both local and global relations. The global geometric arrangement of the data points is non-linear, and some large model variances, observed in a linear model, is due to projections of data clusters that are not one-to-one (the projection function is not injective) – and the variance caused by this noninjective maps are reduced by a projection onto a non-linear manifold. Additionally, capturing the global non-linear geometry of the data serves as a good starting point to decompose the problem into subdomains – identified by clusters – which can be modeled individually. Thus, in some instances, it is best to check for, and model, any global non-linear data set organization as a first step to arriving at a parsimonious model with minimal variance. In addition to its research content, this paper is also intended as a tutorial on how to approach surrogate data modeling of water quality indicators from a manifold learning perspective.

2. METHODS

2.1. De Soto, Kansas: study area and high-frequency USGS data

The USGS station at De Soto, Kansas (Station ID: 06892350; Lat: 38.982, Lon: -94.964) is approximately 32 km downstream from Lawrence, KS, and has a large set of *in situ* water quality sensors including turbidity, chlorophyll-a, and nitrate plus nitrides ($NO_3 + NO_2$) dating back to 2014 and earlier. When we refer to nitrate, nitride, and chlorophyll in this paper, we are actually referencing the quantity measured, ($NO_3 + NO_2$) and chlorophyll-a concentration. USGS measurements are recorded every 15 min. Data from 2014 to 31 March 2023 had discharges in the range of 10^3 to 10^5 ft^3/s, turbidity between 0 and 1,500 FNU, chlorophyll-a concentrations between 0 and 200 μg/L, and nitrate plus nitride concentrations between 0 and 7 mg/L.

2.2. Comment on remote sensing application and variable selection

Given the temporally dense time series record, it is natural to consider the state dependent auto-regressive (i.e., 'dynamic') models for target variables as was recently done to predict nitrate concentrations (Di Nunno *et al.* 2022). However, our ultimate aim in this paper is to develop a model that can be used with satellite data. For that, we need a (non-state-dependent, i.e., 'static') response surface model. Satellite data have extensive spatial coverage but intermittent temporal sampling due to variable atmospheric conditions (e.g., clouds), and sparse sampling (typically with less than daily overpasses). Thus, in this study, we only considered time-independent models with a choice of input variables determined by what is readily available from water quality satellite remote sensing products. The application of the surrogate model developed in this paper is described in a companion paper using data from the satellite sensor Sentinel-2 (Tufillaro *et al.* 2024).

2.3. Data preprocessing: subsetting and normalization

The *in situ* USGS data at De Soto, KS, have temporal data gaps that can last from hours to months. For this study, we aggregated the data into gap-less bundles. Starting with the full data record from 1 January 2014 till 31 March 2023, we first select a set of variables for the surrogate model, typically turbidity, chlorophyll concentration, and discharge as the input variables, and nitrate concentration as the target variable. If gaps in this data are less than 3 h (12 samples) then the gaps are filled in by linear interpolation. The resulting data record is then broken into gapless subsets. All subsets with a gapless record length of less than 2 days (192 samples) are deleted. The remaining records are averaged into 6-h intervals (24 samples), and this final collection of gapless data records is used for

further analysis and modeling. The resulting data set contains 52 gapless data subsets with record lengths varying from 32 (≈2 days) to 2916 (≈182 days). The total data set length is 9,279 points, or (≈6.3 years).

Since the quantities are positive, each time series is normalized between [0, 1]. Where utilized, a seasonal variation variable is defined as

$$s(N) = \left(1 + \sin\left((2 * \pi * N/365) + \frac{3}{2}\pi\right)\right)/2$$

which varies sinusoidally with a minimum at 1 January ($N = 1$) and maximum at day of year $N = 183$. This variable provides a rough proxy for seasonal variations of temperature, sunlight, and seasonal fertilizer applications.

2.4. Data visualization

Our choice of input variables is turbidity, chlorophyll, and discharge (the domain), and nitrate (the co-domain or range). To initially visualize the data we plot each point where the horizontal axes are turbidity and chlorophyll, and the vertical axis tracks nitrate, with discharge indicated by the color scale on the right of Figure 1. Plotting a multivariate time series in a multidimensional space is called an 'embedding' of the data, and depending on the specific choice of input and output variables, results in different multidimensional views. Examining the data in this particular three-dimensional embedding we observe that generally (i) low nitrate values occur at high chlorophyll, (ii) high nitrate values occur at high turbidity and typically high discharge, (iii) a prediction based on a response surface near the origin is highly ambiguous since there is a high variance in the co-domain response for nearby points in the domain. Observations (i) and (ii) are sensible since high chlorophyll is associated with high assimilation of nitrate by phytoplankton, and high turbidity could be associated with runoff presumably rich in nitrate. The wider variance in the range near the origin can be described mathematically by a observing that the mapping appears multi-valued, or 'noninjective'.

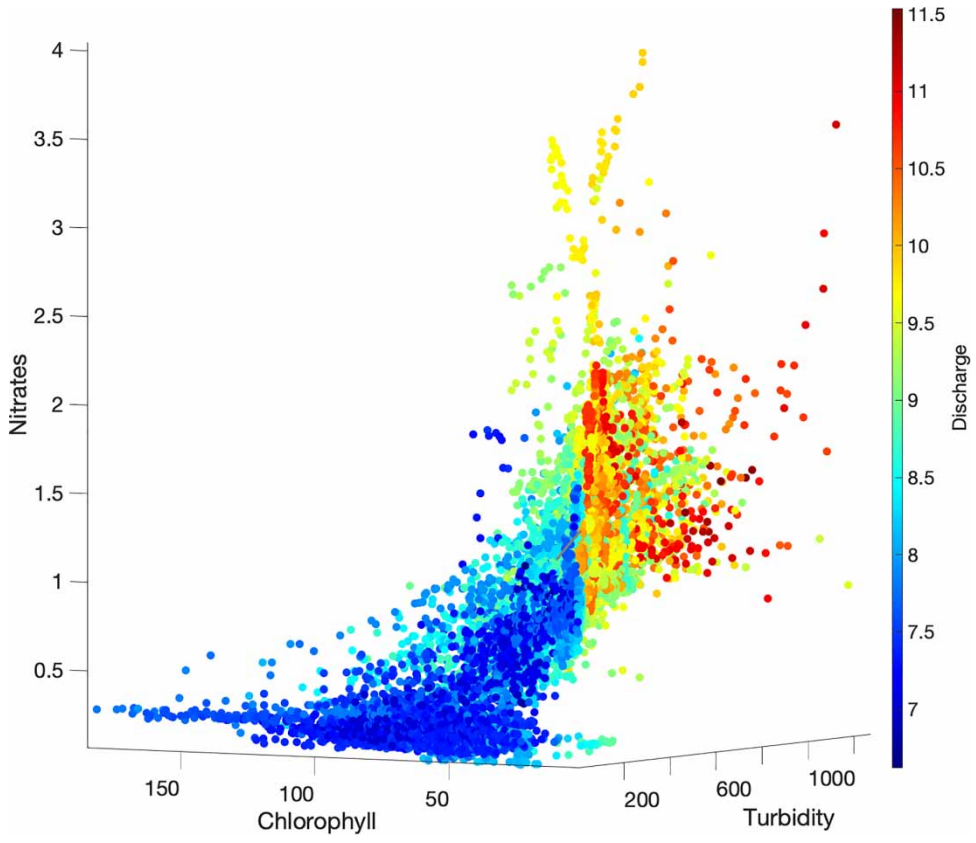

Figure 1 | USGS gauge on the Kansas River at Desoto, KS (06892350) data showing show correlation between turbidity (FNU), chlorophyll-a (μg/L), and nitrate (mg/L).

2.5. Identification of data geometry with biogeochemical processes

To help spread out the data, particularly near the origin, and to more closely approximate a normal distribution, it is common to apply a log transform (in this instance, the natural log is used). Figure 2 shows the log-transformed data projected on a two-dimensional plane (turbidity and chlorophyll) with the nitrate values now represented by a point's color. The log plot (along with plots of the histograms) shows a clustering of the data into two bundles, high chlorophyll, low nitrate (blue region), and high turbidity, high nitrate (red region). Furthermore, Figure 3 shows a similar plot, but with color indicating the season. These two cluster plots paint the following picture. In the winter (near the origin in Figure 3), when low temperatures inhibit phytoplankton growth, there is a moderate level of nitrate, which we assume is mostly moderated by the sediment concentration (as indicated by turbidity). During the first spring bloom, when the temperature reaches 15 °C, chlorophyll levels rise and phytoplankton begin to assimilate nitrate (Bruns *et al.* 2022). As the spring, summer, and fall progress a sequence of blooms oscillates between low and high nitrate states swinging back and forth in the upper right half of the plane in Figure 3, but generally avoiding the moderated nitrate states (low chlorophyll and turbidity)

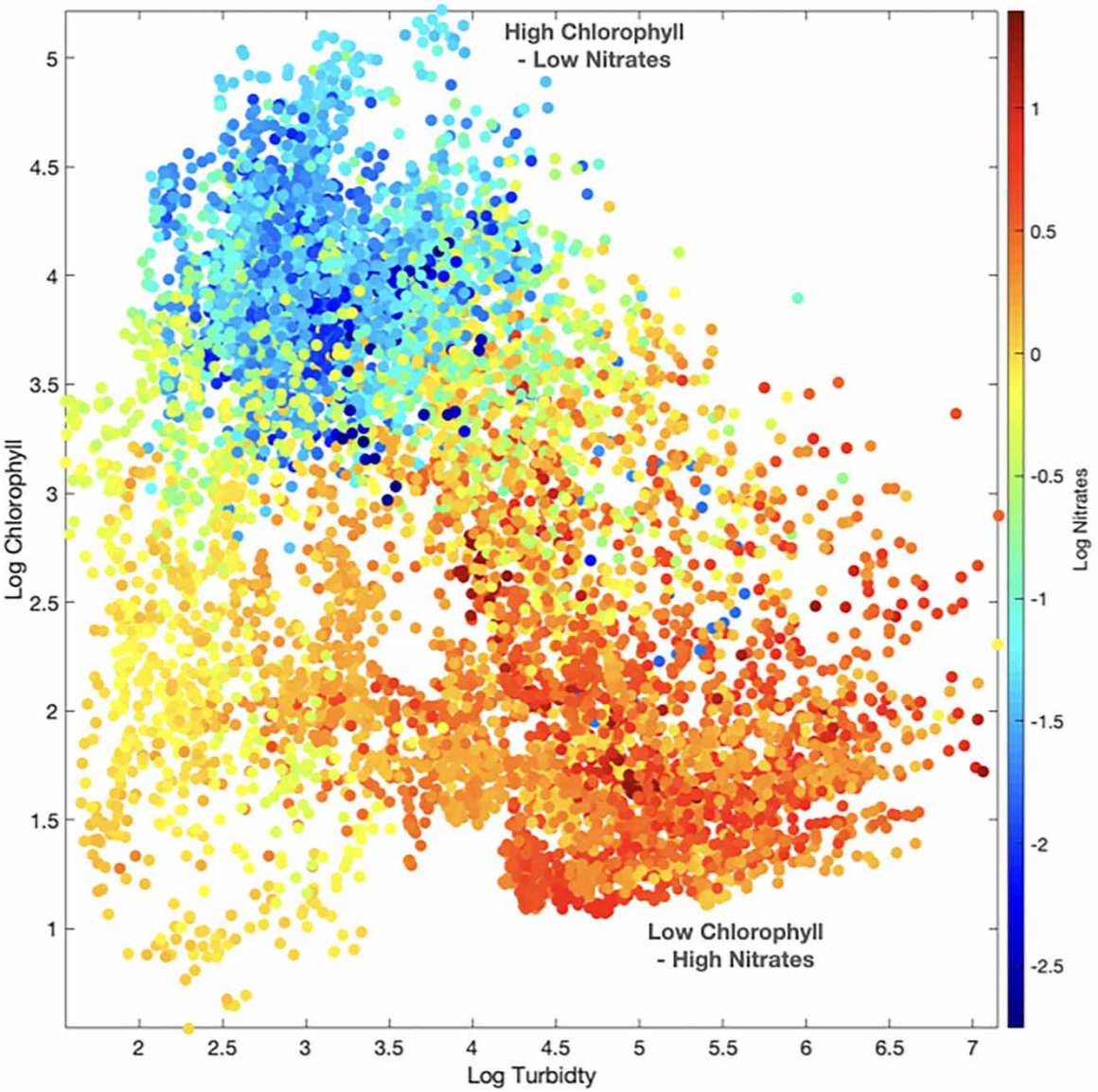

Figure 2 | Log transform of USGS De Soto, KS gauge data showing clustering of nitrate (mg/L) values about high chlorophyll-a concentrations (μg/L) (low nitrate concentrations due to phytoplankton assimilation) and high turbidity (FNU) values. The color of the points indicates the nitrate concentration. The Log transform spreads out the data away from the origin to help reveal the data clustering.

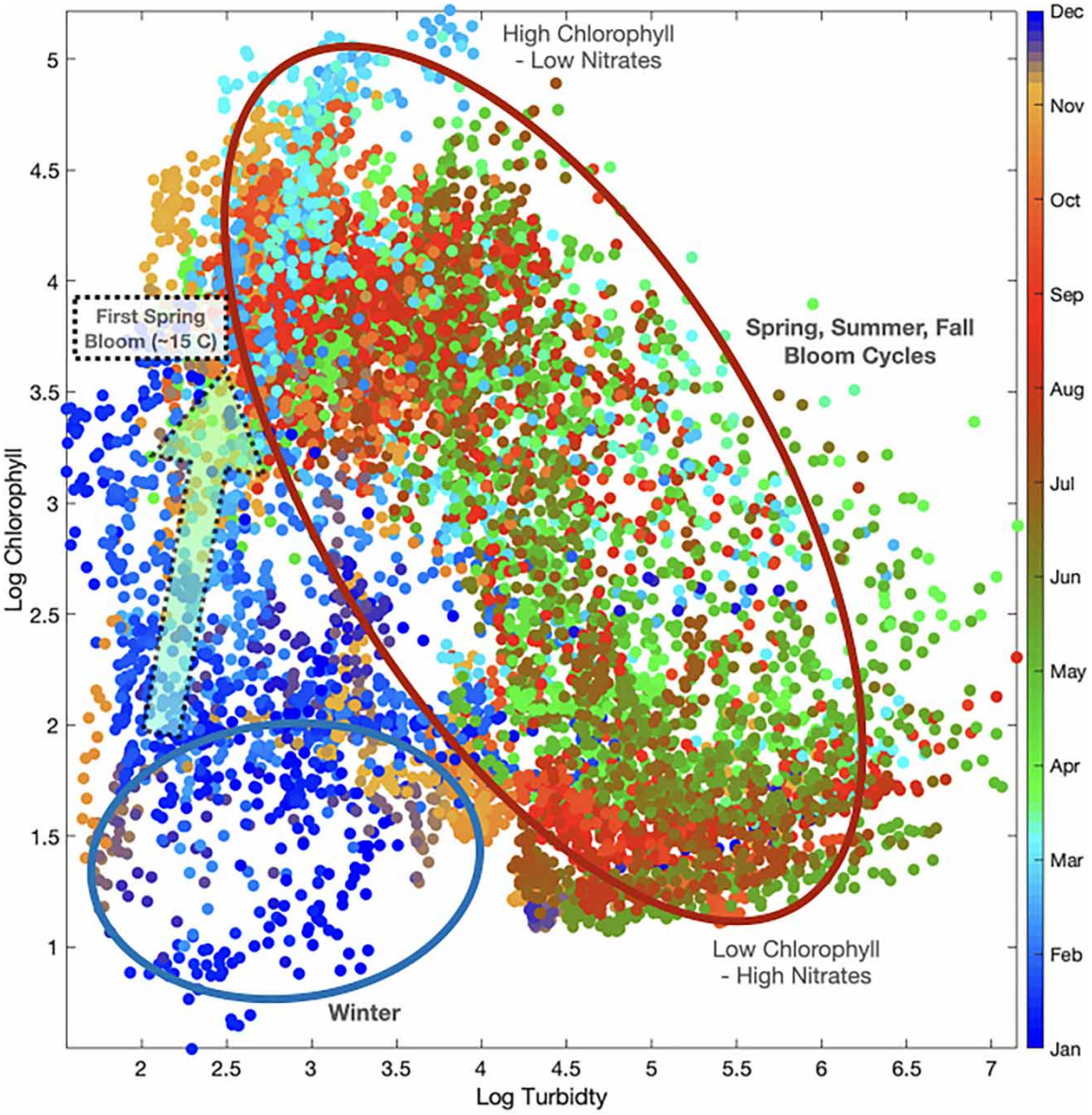

Figure 3 | Log transform of USGS De Soto, KS gauge data showing bloom and best cycle. The color of the points indicates season. Comparison with Figure 2 reveals that moderate nitrate concentrations occur in the winter with a sharp transition to low nitrate values during the early spring when the first phytoplankton bloom occurs (as indicated by high chlorophyll-a concentrations), and then a series of bloom and bust cycles occur during the spring, summer, and fall.

characteristic of winter nitrate levels. This overall seasonal bloom pattern is sometimes described as hysteresis (Burns *et al.* 2019).

It is also observed that the 'variance' of a model prediction can be expected to change by considering different projections of a co-domain into the domain. For instance, in Figure 2, if we consider a model of nitrate using only turbidity as an input variable, then any model prediction will have a high variance for moderate values of turbidity (≈2–4 log(FNU)) since the projection of a point onto the horizontal axis (turbidity) cannot distinguish data points occurring during the winter nitrate state (yellow-orange cluster in Figure 2) and a spring bloom (blue-cyan cluster in Figure 2), so points close together in the domain can take on a wide range of nitrate values simply due to the geometric projection resulting from this choice of embedding. Informally, we say the model embedding has low injectivity to indicate that a partial source of variance is geometric, and not intrinsic data noise.

There are two ways to improve the low injectivity – increase the input dimension, and consider alternative (possibly non-linear) embeddings or projections that better capture the shape of the data. Generally increasing the dimension helps, but also leads to an (exponentially) less dense sampling of the model space – the so-called

curse of dimensionality. The search for non-linear low-dimensional models with lower variance (informally, higher injectivity) is an aim of manifold learning.

Figure 4 shows a three-dimensional plot of the data with the domain again as turbidity and chlorophyll, and the co-domain as nitrate. A linear surrogate the data model is shown as the response surface (a one-to-one, i.e., injective map of the input to output space) is also shown which is computed by standard surface fitting optimization methods (Seber & Wild 2003). Note that because the data is skewed into two separate clusters (a histogram of the data is bimodal, and the data is not normally distributed in the log space), the linear plane approximation misses capturing the obvious non-linear 'shape' of the data. That is, the plane underestimates moderate values of nitrate, and also misses the mean of the data in each cluster – precisely because it is optimizing the mean of all the data points (and implicitly assumes a normal distribution during optimization).

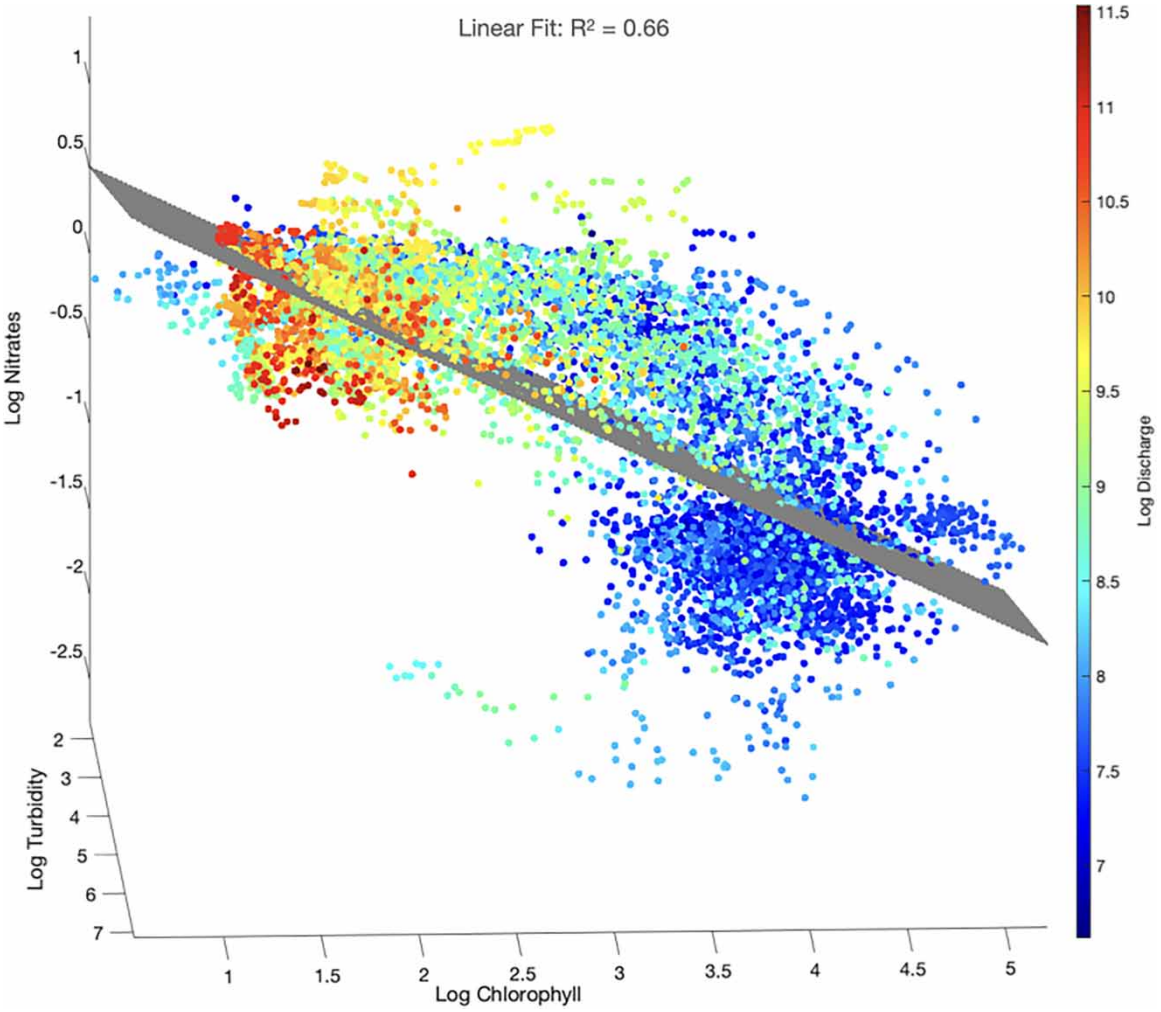

Figure 4 | The non-linear structure of the nitrate data clustering is revealed in a three-dimensional plot of the Log transform. A sharp jump in nitrate concentrations is seen so that the best-fitting planar (linear) fit does not pass through the local mean of the data. A better fit to the data is achieved if we use a non-linear surface that better captures the 'jump' behavior.

The linear surrogate data model has a high variance in part because it is projecting the data onto a linear subspace. If we are able to create a surface (a two-dimensional non-linear manifold, or non-linear subspace) that goes through the mean of the data defined locally, than we can reduce the variance of the surrogate model in two ways: first, by passing more closely through the local mean, we can often remove ambiguities arising from the overlap of projections passing through multiple data clusters, and second, models built from projecting onto this non-linear surface can be closer to a normal distribution. Informally, we refer to this as a mean field model.

The non-linear shape of the data reflects two processes moderating the nitrate concentration. First, (in log space) the nitrate concentration varies approximately linearly with turbidity (and sediment concentration), but non-linearly with phytoplankton assimilation. That is, the sharp non-linear step in the shape of the data is associated with the increase (or decrease) of nitrate assimilation by phytoplankton (see Figure 4 with chlorophyll values between 3.5 and 4.5). Thus, linear surrogate models are adequate for winter nitrate variations but are less than ideal for the spring, summer, and fall, when phytoplankton assimilation kicks in.

2.6. aPL model of data geometry

To include the shape of the data in the modeling process we consider the data from a manifold learning point of view. Roughly, the aim of manifold learning is to find a projection, and change of coordinates, which both reduces the dimensionality of the data model, and also, in a local mean sense, straightens out the data. Mathematically speaking, manifolds need to be 'smooth' objects, which means all their derivatives exists – a (two–dimensional) sphere is smooth, but a two-dimensional cone is not smooth and hence is not a manifold since, at its tip, a local derivative is not well defined. Here we describe the construction of a non-linear surface which is smooth (so technically a manifold) but numerically stiff. It is essentially a spline, but spline constructions are typically not smooth surfaces (i.e., there are discontinuities in some of their derivatives).

To capture the data's shape we start by partitioning the domain into two regions which separate the clusters of high and low nitrate regimes identified in Figure 2. Figure 5 illustrates how to create a domain indicator function which is close to 1 (respectively 0) on region d_1, and 0 (respectively 1) on region d_2:

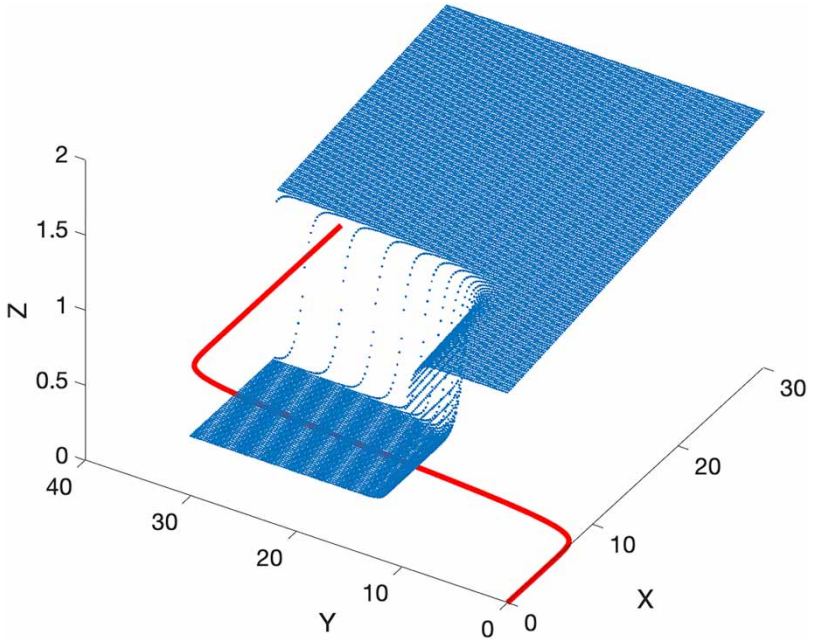

Figure 5 | A simple model that captures the non-linear 'jump' (bloom-bust cycle, Figures 3 and 4) consists of two planes (around each cluster), and is continuously connected with a non-linear function (a hyperbolic tangent in this instance). A non-linear fit to this aPL model better captures the (mean) non-linear structure in the data with a model of only just a few parameters (11 parameters in this example). The red curve in the X–Y plane marks the separated domains of the planar solution about each cluster.

$$d_1(x,y) = \frac{1}{2}\left(1 + \tanh\left(\omega_2 \cdot \left(y - \frac{\beta}{2}(1 + \tanh(\omega_1 \cdot (x - \alpha_1)))\right) - \alpha_2\right)\right)$$

$$d_2(x,y) = \frac{1}{2}\left(1 - \tanh\left(\omega_2 \cdot \left(y - \frac{\beta}{2}(1 + \tanh(\omega_1 \cdot (x - \alpha_1)))\right) - \alpha_2\right)\right)$$

where the five constants $(\alpha_i, \beta, \omega_i)$ with $i = \{1, 2\}$ are parameters that control the location and steepness of the cross-over from region d_1 to d_2. Each region is now modeled by another function which multiplies the domain

indicator. Here we consider only linear functions of the form:

$$f_1(x, y) = (a_1x + b_1y + c_1) \cdot d_1(x, y)$$
$$f_2(x, y) = (a_2x + b_2y + c_2) \cdot d_2(x, y)$$

which we describe as aPL. The constants a_i, b_i, c_i with $i = \{1, 2\}$ define the slope and offset of the linear approximation on each subdomain. The use of the *tanh* function is a convenient choice since it can be adjusted to approximate the basic outline of the data. Other selections of domain indicator functions and modeling functions are possible. The prefix 'almost' is a reminder that, because of the non-linear cross-over function, the model is not exactly piece-wise linear. The model contains 11 parameters, so by design, it won't overfit the data. The behavior outside the domain of the training date extrapolates nicely as well.

The constants are estimated using non-linear optimization, specifically the *fitnlm* function in MATLAB (Seber & Wild 2003). Non-linear optimization requires a choice of starting initial conditions, and a good way to pick those is to estimate an initial set of constants for the model with a *tanh* curve separating the two clusters, and an approximate mean value for the response surface in each cluster.

Two views of a fit to the data of the aPL model are shown in Figures 6 and 7. Figure 6 shows how at low chlorophyll values, the log of the turbidity approximately increase linearly. In contrast, Figure 7 shows the rapid non-

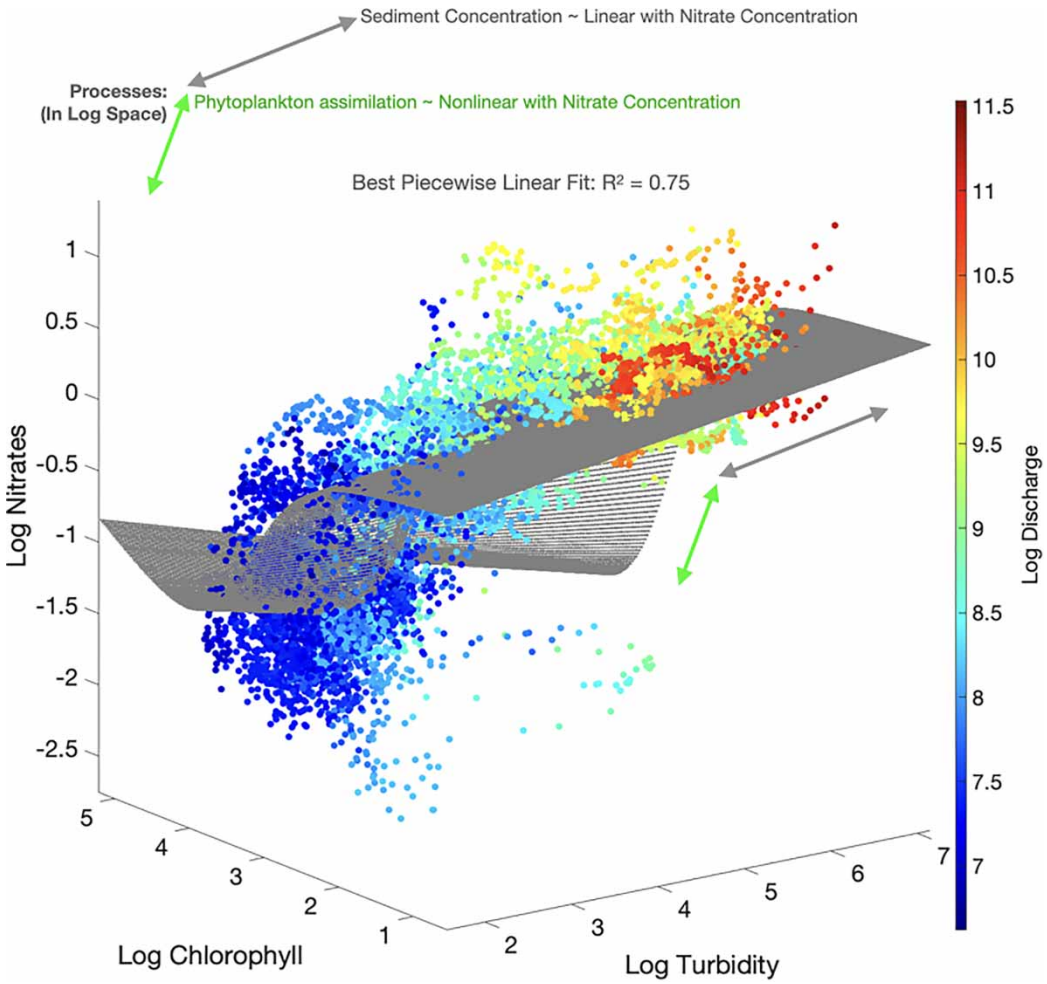

Figure 6 | A non-linear optimization of the aPL model to the data shows that the (Log) of nitrate concentrations at low chlorophyll values is approximately linearly proportional to turbidity, and the transition to the high chlorophyll state is non-linear. The color of the points is proportional to the discharge which also indicates that low chlorophyll states typically occur at high discharge values, and low chlorophyll concentrations typically occur low discharge values – presumably in stiller water. This suggest that river 'flushing' from control structures (reservoirs) is an effective method to dissipate blooms. Model values: $\beta = 0.5961$, $\alpha_1 = 3.8200$, $\alpha_2 = 7.7535$, $\omega_1 = 4.0395$, $\omega_2 = 2.3029$, $a_1 = -0.0131$, $b_1 = 0.2894$, $c_1 = -2.5829$, $a_2 = 0.0922$, $b_2 = -0.1125$, $c_2 = 0.0985$.

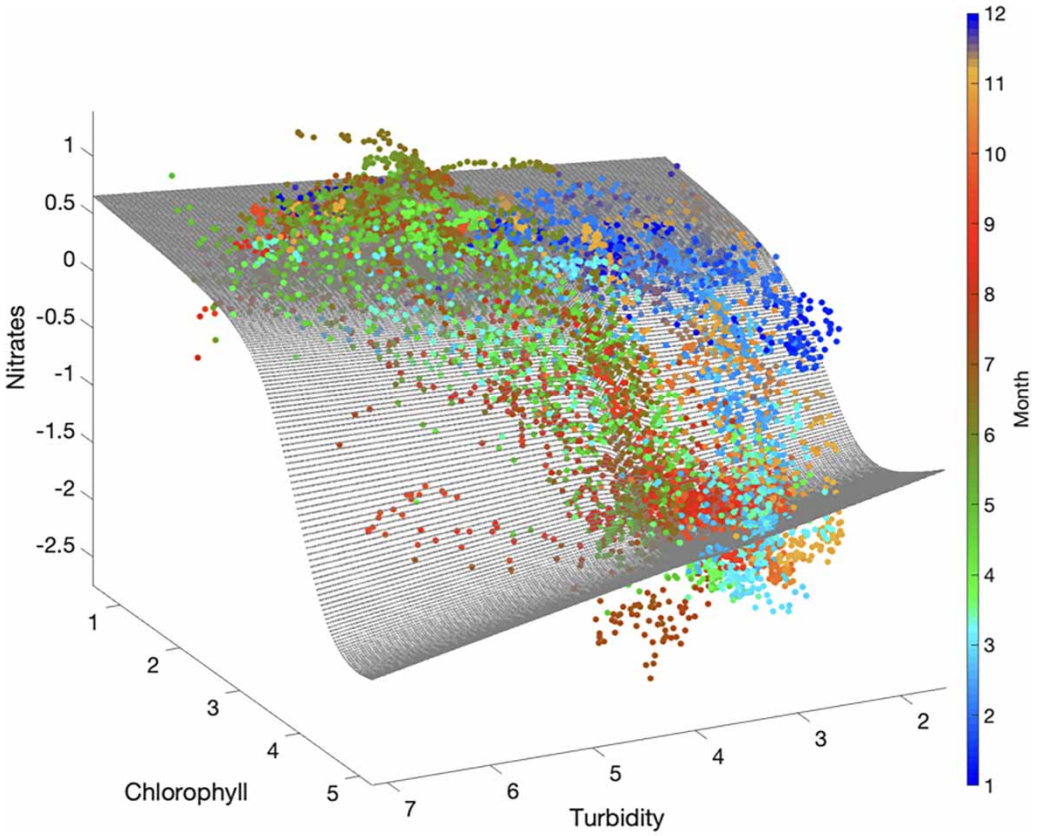

Figure 7 | Rotation of Figure 6 provides an alternative view of the data points and the model fit. The time of year is indicated by the point color showing that the winter months (blue) have moderate nitrate concentrations which then typically rapidly decrease in the early spring (cyan) followed by bloom and bust cycles occur the rest of the year (green and red hues).

linear switching of nitrate concentration between the high and low nitrate states. The aPL model also highlights the hole between the first winter-spring bloom, and the blooms occurring the rest of the year. The shape of the data is not only nonlinear, but also has one hole, topologically speaking.

2.7. Regression and error metrics

Two metrics are used to evaluate the model quality. The *coefficient of determination*, R^2, is an indicator of the how much of the variance in the dependent variable is explained by the (multiple) independent variables. This is provided for the linear surface fits in Figure 4 and the aPL surface fit in Figure 6. A value of $R^2 = 0$ indicates that the model is no better than using the mean value for estimation; a $R^2 = 1$ indicates perfect correlation. Since we are interested in how well the target variable of nitrate is predicted, we also present a *one-to-one* line ($y_{obs} = 1 \cdot y_{model} + 0$) and its single variable linear *goodness-of-fit* metric (r^2), the difference between the model prediction value (horizontal axis) and the measured value (vertical axis). The goodness-of-fit is a useful metric to consider when estimating daily or seasonal nitrate loads. Neither metric indicates causation, only a correlation between the independent and dependent variables.

3. RESULTS AND DISCUSSION

Figure 8 (corresponding with Figure 4) shows the one-to-one line and goodness-of-fit of the linear nitrate model with inputs consisting of turbidity and chlorophyll-a concentration. Notice that around the low chlorophyll cluster (with the nitrate values less than 0.5 mg/L), the one-to-one line and linear model prediction data are biased for low values of nitrate This bias is also indicated by the linear model slope of $0.81 < 1$. Again, the large variance in this parameter regime is, in part, due to points from the two different data clusters being mixed in the data projection onto a linear subspace. In contrast Figure 9 (corresponding to Figure 6), the (almost) Piece-wise linear one-to-one line, and linear model fit, results in a slope of $\approx$1.0 indicating that the model passes very closely

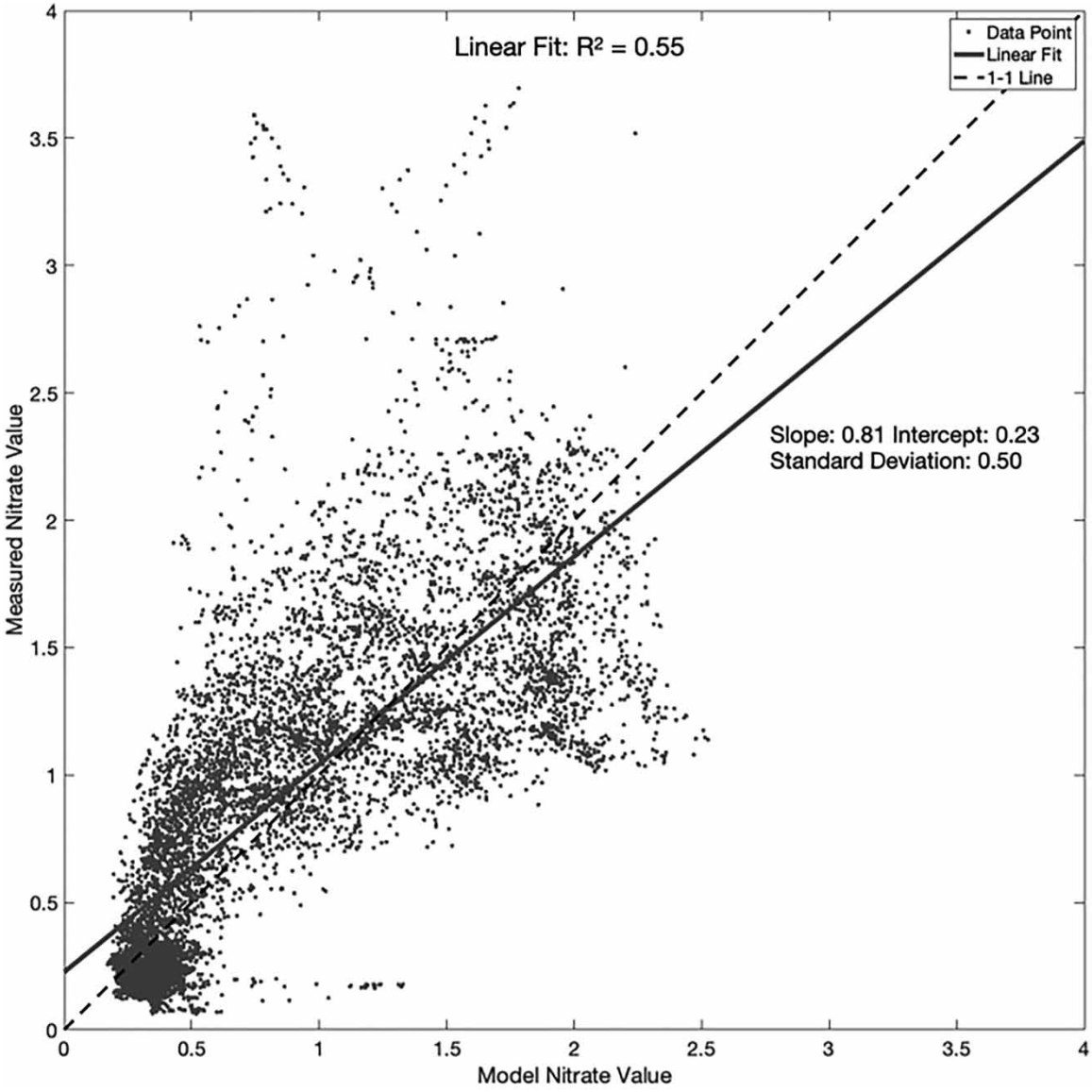

Figure 8 | One-to-one linear (bold) and linear fit (dashed) of the predicted (horizontal axis) and measured nitrate concentrations (vertical axis) based on the linear (planar) surface fit (Figure 4). The slope 0.81 < 1 indicates a bias in the model due to the fact that the linear fit does not pass though the local mean of the data. The quality of the fit is gauged by the r-squared value of 0.55.

though the mean for the bulk of the data, that is, the aPL model produces predictions that pass through the mean of each cluster simultaneously.

The increase in model quality from a linear model to a non-linear (i.e., almost piece-wise linear model) is indicated by the metrics: The R^2 increase from 0.66 to 0.75, and the goodness of fit (r^2) increase from 0.55 to 0.68.

The best test of the surrogate nitrate model is the *out-of-sample* predictions. Figure 10 shows a time series of the results for a training set to early 2020, the model prediction for the training data is indicated in blue. The red data points indicate the model prediction for the out of sample data from early 2020 through to March 2023. The goodness of fit parameter is $r^2 = 0.65$ in each time series. That the values are identical is coincidence, but the fact that the modeled times series looks similar before and after 2020, and that the r^2 metrics are close, indicates that model performs well on unseen future nitrate states.

Also, Figure 10 shows that the model does not predict extreme events well. The model predictions on each cluster are limited by the linear assumption of the model on each domain, and the variability permitted by the parameter fit to this locally planar model. To handle extreme events we next develop a model specific to the high nitrate (domain *d*2) cluster. A good approach to modeling data on each cluster is the use of additional variables such as discharge, or a seasonal variable, to further disentangle the observed nitrate values.

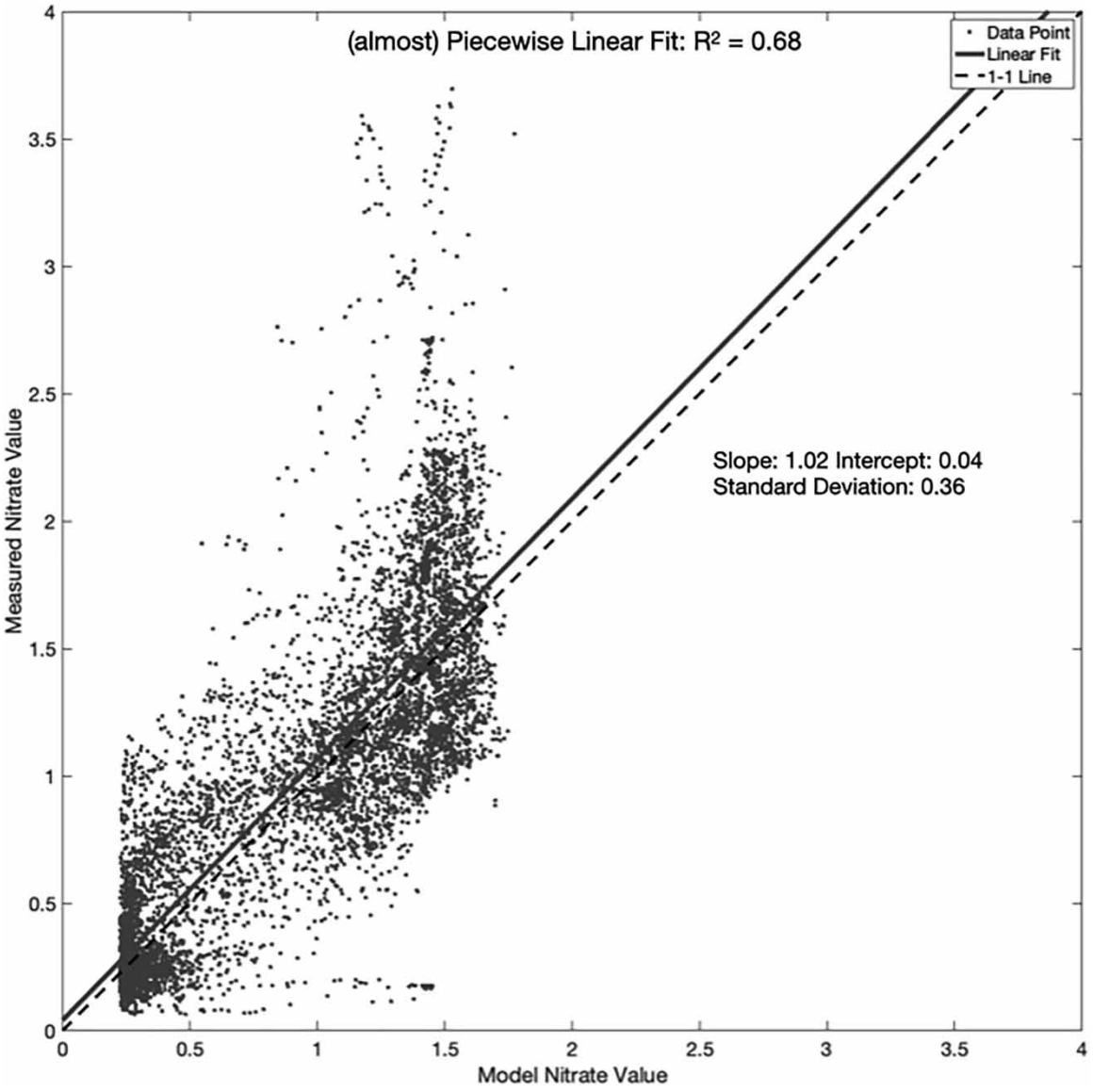

Figure 9 | One-to-one linear (bold) and linear fit (dashed) of the predicted (horizontal axis) and measured nitrate concentrations (vertical axis) based on the aPL surface fit (Figure 6). The slope ≈ 1 indicates minimal bias. The quality of the fit is gauged by the r-squared value of 0.68. See Figure 10 for the time series comparison.

A probabilistic model, such as a Gaussian Process model or Bayesian Markov chain Monte Carlo estimation, is likely the best approach to predicting the nitrate values on each cluster (Gershenfeld 1999). Here we provide a more elementary discussion in keeping with the response surface model already developed for the local mean response described by the aPL model. We are interested in capturing high concentration pulses of nitrate. Therefore we augment the response surface on the high nitrate cluster by computing a mean value of nitrate as a function of the (mean value) of discharge and season as indicated by a 'day of year' (N) variable capturing seasonality. On the high nitrate cluster (as defined by $4 <$ Log(Turbidity) < 6.6, and $1.2 <$ Log(Chlorophyll) < 2.5), we first subtract the mean value predicted by the aPL model, and then divide up the discharge-seasonality domain variables into $N \times N$ bins ($N = 7$ in Figure 11). In each bin we compute the mean values of the input and response variables. This produces a set of points estimating the average nitrate values in the high nitrate cluster as a function of the local means for discharge and season. The scattered point data of nitrate values is then used to estimate a surface using radial basis functions (RBF). The seasonality domain is periodic, so we pin the seasonality boundaries (N between 1 and 365), and the discharge boundaries, to the mean values predicted by the aPL model. In the fitting process the 2D Gaussian radial basis functions have a tuning parameter σ which regularizes (controls the smoothness) of the estimated solution (Chirokov 2023). The exact estimate solution is performed with using the RBF

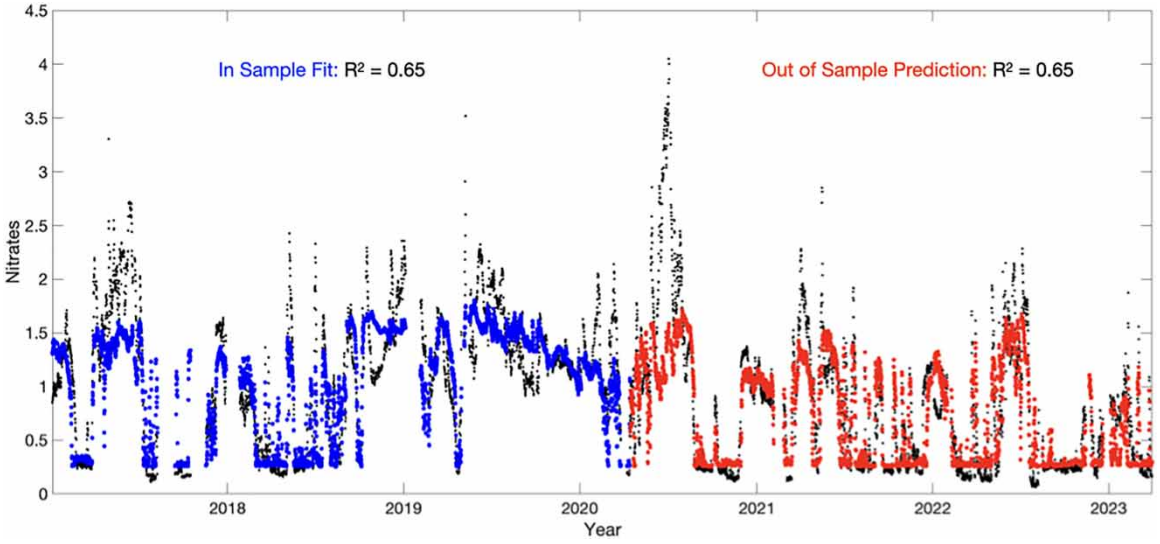

Figure 10 | Out of sample prediction for the aPL nitrate concentration model. The time series to the left (blue) is the fitting data. The time series to the right (red) is the out of sample test – unseen data not from the training data set. The exact same r-squared value is a coincidence, but closely similar r-squared values indicates similar model performance on both the training and the test set.

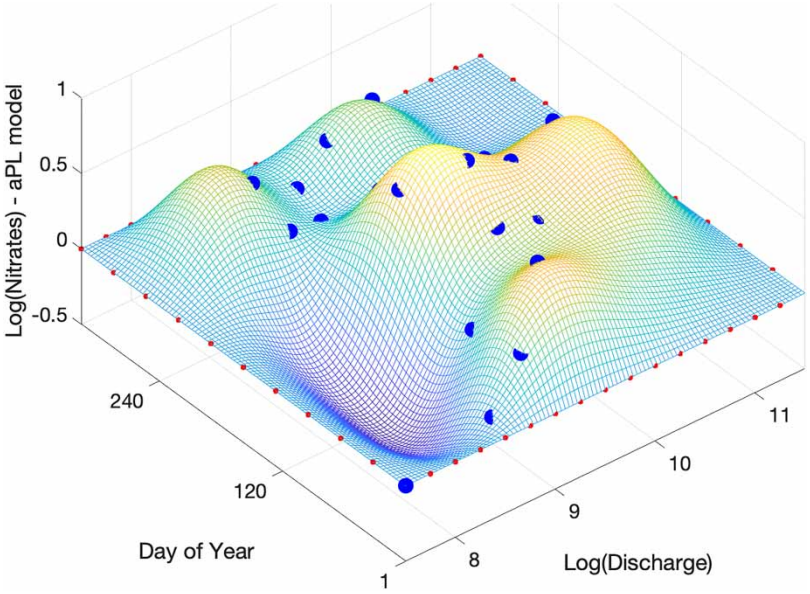

Figure 11 | A Radial Basis Function (RBF) model for the high nitrate cluster. The blue set of training points is computed by averaging over the original data set enclosed by the high nitrate cluster (Figure 6). As illustrated in Figure 12, it captures the extreme high nitrate events better than the original aPL model (Figure 10).

solver for Matlab authored by Chirokov and is sensitive to many parameters – the regularization parameters (σ), the specification of the boundary conditions, the binning choice (N), and so on. However, as illustrated in Figure 11, estimated solutions, for reasonable choices of fitting parameters, produce plausible results.

Figure 11 shows an example surface fit for predicting nitrate in the high nitrate cluster. The fit indicates that the maximum value of nitrate – on average – occurs late spring-early summer with a discharge in the vicinity of 9.5 ($\approx$13,000 ft^3/sec) – that is, higher and lower values of discharge tend to result in lower nitrate values. Further, nitrate values on the high nitrate cluster are suppressed in the early spring at lower discharge rates. The perhaps nonobvious result, suggested by this particular estimation method, is that (log) of discharge values in the range of 9.5–10 ($\approx$13,000–20,000 ft^3/sec), on average, are correlated with higher nitrate concentrations at this specific site.

The time series comparison is shown in Figure 12. The boost in signal provided by the RFB model on the high nitrate cluster visually shows better agreement between the measured data and RBF model for nitrate pulses, and

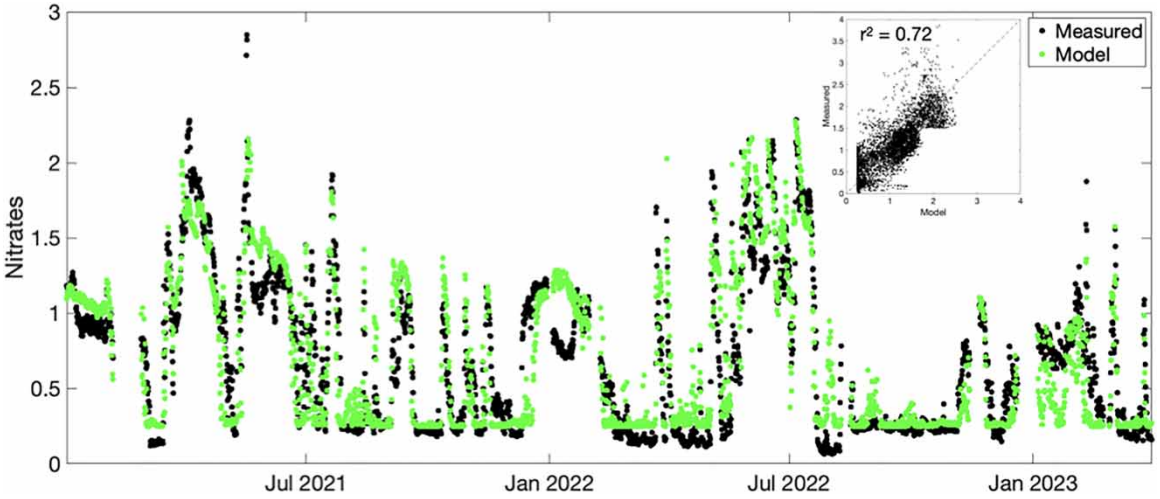

Figure 12 | The time series compares predictions from the aPL model augment with the RBF model for the high nitrate cluster. A visual comparison shows that the augmented model captures the extreme high nitrate evens more accurately and this is reflected in the greater r-squared value of 0.72 shown by the one to one plot presented in the inset.

this is reflected in the modest increase in the value $r^2 = 0.72$. Being able to better estimate these high concentration nitrate events should improve estimates of seasonal nitrate loads based on surrogate data time series models.

4. CONCLUSION

This paper presents a visually rich introduction to surrogate time series methods for estimating nitrate with an emphasis on connecting the geometry observed in the time series data to the underlying biogeochemical processes. The work demonstrates that, during rapid nitrate assimilation by phytoplankton, changes in nitrate concentrations are non-linear relative to variations in turbidity. We present a simple data driven method to model this non-linearity by describing an (almost) Piece-wise linear response surface that provides a mean field approximation to the dynamics of the measured data for nitrate plus nitrite correlations to turbidity and chlorophyll-a concentrations. We argue that one advantage to switching to a coordinate system lying on a mean field approximation surface is that some of the model variance observed in linear models is due to a projections that mix points from different data clusters, and that by considering a projection to a non-linear subspace, the model variance is reduced. The method extends the USGS linear procedures for surrogate data modeling allowing for better approximations for river systems exhibiting algal blooms due to nutrient rich source waters. There is nothing in the model that is unique to the Kansas river compared to similar rivers in the American agricultural midwest. Therefore, the model and visualization procedures illustrated in the Kansas river example should be generally applicable to many medium size rivers in agricultural regions.

DATA AVAILABILITY STATEMENT

All relevant data are available from an online repositoryor repositories (please ensure the DOI/URL has been provided as a submssion item).

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Burns, D. A., Pellerin, B. A., Miller, M. P., Capel, P. D., Tesoriero, A. J. & Duncan, J. M. 2019 Monitoring the riverine pulse: Applying high-frequency nitrate data to advance integrative understanding of biogeochemical and hydrological processes. *Wiley Interdisciplinary Reviews: Water* **6**(4), e1348. doi:10.1002/wat2.1348.

Bruns, N. E., Heffernan, J. B., Ross, M. R. V. & Doyle, M. 2022 A simple metric for predicting the timing of river phytoplankton blooms. *Ecosphere* **13**(12), e4348.

Chirokov, A. 2023 Scattered data interpolation and approximation using radial base functions (https://www.mathworks.com/matlabcentral/fileexchange/10056-scattered-data-interpolation-and-approximation-using-radial-base-functions">https://www.mathworks.com/matlabcentral/fileexchange/10056-scattered-data-interpolation-and-approximation-using-radial-base-functions), MATLAB Central File Exchange. Retrieved May 27, 2023.

Di Nunno, F., Race, M. & Granata, F. 2022 A nonlinear autoregressive exogenous (NARX) model to predict nitrate concentration in rivers. *Environmental Science and Pollution Research* **29**(27), 40623–40642.

Foster, G. M. & Graham, J. L. 2016 *Logistic and linear regression model documentation for statistical relations between continuous real-time and discrete water-quality constituents in the Kansas River, Kansas, July 2012 through June 2015*. U.S. Geological Survey Open-File Report 2016–1040, p. 27. doi:10.3133/ofr20161040.

Gershenfeld, N. A. 1999 *The nature of mathematical modeling*. Cambridge University Press, Cambridge, UK.

Rasmussen, T. J., Ziegler, A. C. & Rasmussen, P. P. 2005 *Estimation of constituent concentrations, densities, loads, and yields in lower Kansas River, northeast Kansas, using regression models and continuous water-quality monitoring, January 2000 through December 2003*. U.S. Geological Survey Scientific Investigations Report 2005–5165.

Rasmussen, P. P., Gray, J. R., Glysson, G. D. & Ziegler, A. C. 2009 Guidelines and procedures for computing time-series suspended-sediment concentrations and loads from in-stream turbidity-sensor and streamflow data: U.S. Geological Survey Techniques and Methods book 3, chap. C4, p. 53.

Seber, G. A. F. & Wild, C. J. 2003 Nonlinear Regression. Wiley-Interscience, Hoboken, NJ.

Tufillaro, N., Piazza, B. P., Reddy, S., Baustian, J., Sousa, D., Grötsch, P., Lalović, I., De Moitié, S. & Zurita, O. 2024 Linking optical data and nitrates in the Lower Mississippi River to enable satellite-based monitoring of nutrient reduction goals. *Ecohydrology*, e2631. https://doi.org/10.1002/eco.2631.

Van Der Maaten, L., Postma, E. & Van den Herik, J. 2009 Dimensionality reduction: A comparative review. *Journal of Machine Learning Research* **10**, 13.

Williams, T. J. 2021 *Linear regression model documentation and updates for computing water-quality constituent concentrations or densities using continuous real-time water-quality data for the Kansas River, Kansas, July 2012 through September 2019* U.S. Geological Survey Open-File Report 2021–1018, p. 18. https://doi.org/10.3133/ofr20211018">https://doi.org/10.3133/ofr20211018 (see Appendix 18 for USGS linear surrogate models of Total Nitrogen at the De Soto, KS site).

Yang, G. & Moyer, D. L. 2020 Estimation of nonlinear water-quality trends in high-frequency monitoring data. *The Science of the Total Environment* **715**, 136686. PMID: 32032984. doi:10.1016/j.scitotenv.2020.136686.

First received 23 August 2023; accepted in revised form 13 February 2024. Available online 21 March 2024