Revised: 17 January 2024 | Accepted: 28 January 2024

DOI: 10.1002/eco.2631

RESEARCH ARTICLE

WILEY

Linking optical data and nitrates in the Lower Mississippi River to enable satellite-based monitoring of nutrient reduction goals

Nicholas Tufillaro[1] | Bryan P. Piazza[2] | Sheila Reddy[3] | Joseph Baustian[2] | Dan Sousa[4] | Philipp Grötsch[5] | Ivan Lalović[6] | Sara De Moitié[7] | Omar Zurita[8]

[1]Gybe, Corvallis, Oregon, USA

[2]The Nature Conservancy, Baton Rouge, Louisiana, USA

[3]Chief Strategy Office and Global Science, The Nature Conservancy, Durham, North Carolina, USA

[4]Department of Geography, San Diego State University, San Diego, California, USA

[5]Gybe, Berlin, Germany

[6]Gybe, Portland, Oregon, USA

[7]Gybe, Peniche, Portugal

[8]Gybe, San Diego, California, USA

Correspondence
Nicholas Tufillaro, Gybe, PO Box 1028, Corvallis, OR 97333, USA.
Email: nick@gybe.eco

Funding information
Enterprise Rent-A-Car Foundation

Abstract

Hypoxic zones and associated nitrate pollution from farms, cities and industrial facilities is driving declines in water quality that affect ecosystems, economies and human health in major rivers and coastal areas worldwide. In the Mississippi River, the United States Environmental Protection Agency set a goal of reducing nitrogen loading 20% by 2025, but estimating progress towards this goal is difficult because data from in-stream gauges and laboratory samples are too sparse. Satellites have the potential to provide sufficient data across the Mississippi River, if a key methodological challenge can be overcome. Satellites provide data from visible light, but nitrates are only observable with ultraviolet light. We address this methodological challenge by using a two-step surrogate modelling procedure to link optical data and nitrates in the Lower Mississippi River. First, we correlate in situ nitrate measurements to common water quality parameters, particularly turbidity and chlorophyll, using data from water sensors installed at Baton Rouge, Louisiana, USA, and a long-term dataset from Louisiana State University. Second, we correlate these water quality data to satellite estimates of water quality parameters. We found a correlation between these water quality parameters and nitrate concentrations, as indicated by a coefficient of determination, when the relationship was viewed in nonlinear parameter space. The spatial extent of the correlation was tested with an upstream nitrate sensor 140 km north of the estimation location. These results provide proof of concept that we can develop models that use satellite data to provide large-scale monitoring of nitrates across the Mississippi River Basin and other impaired rivers, globally.

KEYWORDS

Baton Rouge, hyperspectral, monitoring, nitrates, remote sensing, surrogate modelling

1 | INTRODUCTION

A major threat affecting coastal and marine ecosystems worldwide is the proliferation of low oxygen, hypoxic zones. These areas, defined as having bottom waters with dissolved oxygen values <2 mg/L, currently number over 400 globally (Diaz & Rosenberg, 2008). Hypoxia is primarily caused by excess nutrients (nitrogen and phosphorus) that originate from farms, cities and industries, often in inland watersheds, far removed from the coastal areas they impact. These excess nutrients drive spatially extensive growths of algae in the coastal ocean that, when they die and sink to the bottom, rapidly decompose and deplete the oxygen in the bottom water. Coastal hypoxia, which typically happens during the summer months, is further exacerbated by water column stratification, which precludes mixing of bottom water with oxygen-rich surface water. In addition to the local effects of excess nutrient loading into waterways, coastal hypoxia has been linked to

https://doi.org/10.1002/eco.2631

reductions in aquatic biodiversity, fishery valuation and tourism (Rabotyagov, Kling, et al., 2014; Smith et al., 2017). A key challenge in achieving water quality goals aimed at reducing dead zones is a lack of data for monitoring changes in nutrient concentrations in rivers.

In the United States (US), the US Environmental Protection Agency (EPA) set a goal to reduce excess nutrient loading to the Gulf of Mexico (GOM) by 20% by 2025 (Greenhalgh & Sauer, 2003; Rabotyagov, Kling, et al., 2014). The GOM Dead Zone is the second largest in the world (5-year average km^2, range 5000–22,000 km^2 Rabalais, 2011; Rabalais et al., 2002; https://www.epa.gov/ms-htf/northern-gulf-mexico-hypoxic-zone).

Driven by flows from the Mississippi River (MR), which drains the 3.2 million km^2 Mississippi River Basin (MRB, the fourth largest river basin in the world), the GOM Dead Zone, typically occurs off the Louisiana coast during the summer. In 1997, the US EPA established the MR/GOM Hypoxia Task Force (Hypoxia Task Force [HTF]) to understand the causes and effects of GOM hypoxia and coordinate efforts to reduce the size, severity and duration of the hypoxic zone.

To do this, the HTF promoted the formation of research programmes, partnerships and legislation, aimed at meeting these goals. In 2001, the HTF published the first Hypoxia Action Plan, a national strategy to reduce the frequency, duration, size and intensity of the GOM Dead Zone. This strategy, which was updated in 2008 and 2015, provides coastal, within-basin and quality-of-life goals. The coastal goal is to reduce the 5-year running average areal extent of the GOM hypoxic zone to less than 5000 km^2 by 2035 with an interim 20% reduction of nitrogen and phosphorus loading by 2025. The within-basin goal is to restore and protect the waters of the 31 states and tribal lands within the MRB through nutrient and sediment reduction actions to protect public health and aquatic life. The quality-of-life goal seeks to improve communities and economic conditions (agriculture, fisheries and recreation) across the MRB through improved public and private land management and incentives to improve water quality (full policy goals, implementation strategies and action plans can be found at https://www.epa.gov/ms-htf).

These EPA policy goals have driven investment in developing practices to reduce nutrient loading into MRB waterways and measure the effects of these practices. There is evidence that these practices reduce nutrient loads, especially for nitrate (NO_3–), the inorganic form of nitrogen that largely drives the algal lifecycle responsible for GOM hypoxia. While there have been many research and monitoring studies showing the effects and cost-effectiveness of practices designed to reduce nutrients (Bennett et al., 2016; Greenhalgh & Sauer, 2003; McLellan et al., 2015; Rabotyagov, Campbell, et al., 2014), and many others, particularly nitrate, one problem is that current water monitoring to measure nitrate is lacking. Nitrate absorbs light in the ultraviolet (UV) spectrum (200–205 nm Edwards et al., 2001). Therefore, it is typically measured either with in-water UV optical sensors, mounted to continuous water quality gauges, or from discrete water quality samples analysed with UV spectroscopy in a laboratory. Both methods are expensive and time consuming, and as a result, stations are relatively sparse, and large reaches of the MRB remain unmonitored.

Remote sensing using satellites provides the ability to monitor the earth at greater spatial scales by allowing for low-cost, repeated 'virtual gauging' of waterbodies. Remote detection of water constituents with a spectral signature outside of visible light spectrum, that is, a UV and IR spectral signature, is not possible because of the high absorption of water at these wavelengths. One approach to circumvent this issue is to search for correlations to visible spectral features. For instance, estimation of nitrate concentrations in rivers from correlations to other water parameters (e.g., discharge, turbidity and specific conductance) has been developed by the US Geological Survey (USGS) under the framework of surrogate modelling (Rasmussen et al., 2005; Williams, 2021). Predictive models of nitrate concentrations have also been demonstrated (Di Nunno et al., 2022) using nonlinear autoregressive models using only inputs of discharge, dissolved oxygen, specific conductance and water temperature. Previous work using surrogate models focused on correlations between lab sample analysis from water grabs and in-water nutrient gauges. Here, we extend this analysis to include remote sensing data, thereby going from sparse sampling of single points to spatially dense sampling of a region, which is necessary to provide proof of concept for basin-wide monitoring to track progress on the US EPA nutrient reduction goal.

We build on the surrogate modelling approach with a two-step procedure that uses both in situ and satellite data. First, we correlate in situ nitrate measurements to common water quality parameters, in particular turbidity and chlorophyll, with a training dataset formed from an above-water sensor installed for 36 months, in-water sensors operated by the USGS and a long-term dataset from Louisiana State University (LSU). Second, we correlate the local gauge data to satellite estimates of the same water quality parameter from Sentinel-2.

Informally, we refer to the training data models as static, because they only involve function estimation, and we refer to the nonlinear time series analysis methods as dynamic, because (due to feedback terms) they require state estimation. Here, we limit our study to (static) surrogate model estimations of nitrate concentration in rivers because the remote sensing data are both less frequent and have nonuniform sampling (due to issues such as cloud cover), which precludes the straightforward use of dynamic models. Our results, which show a correlation between visible water quality constituents and nitrate concentrations in the Lower MR, are proof of concept for satellite-based monitoring of nitrates. Insights gained through developing this method can help inform the development of a system of models that would use satellite data to remotely monitor nitrates across the MRB, ultimately enabling estimates of progress towards the US EPA's goal of reducing nitrates.

2 | METHODS

2.1 | Study area

We conducted this proof of concept in the Lower MR, which receives drainage water from the entire MRB (Piazza, 2014). This is a highly turbid river environment with stable nitrate concentrations

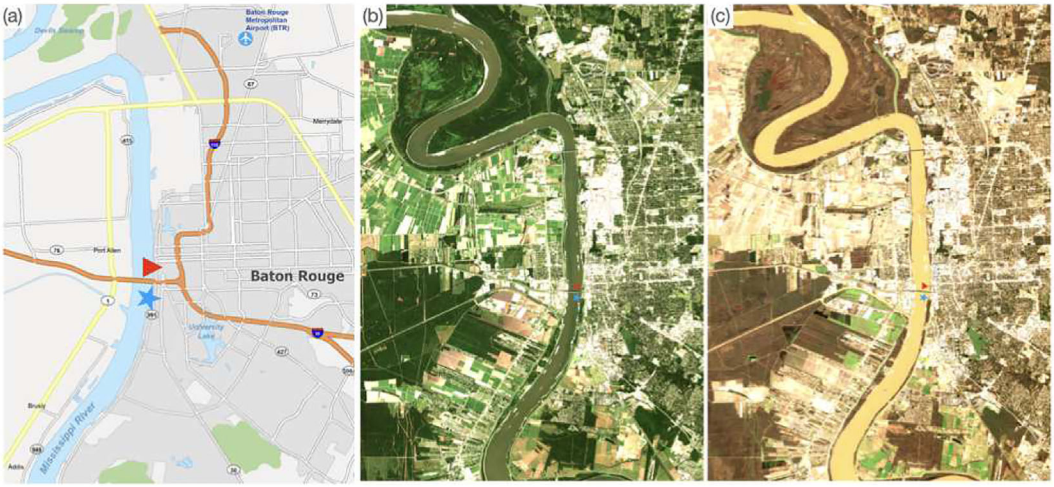

FIGURE 1 (a) Map of Baton Rouge, LA, with markings for the location of The Water Institute of the Gulf (blue star) and the Public Dock (red triangle). (b) Sentinel-2 image of Baton Rouge, LA, on 18 October 2017 with a chlorophyll-a concentration of 34.4 μg/L. (c) Sentinel-2 Image from 31 January 2018 with high turbidity of 91.8 FNU. The images are not colour accurate RGBs and are constructed from the Sentinel-2 bands (0.665, 0.560, 0.443) nm; thus, the sediment appears more yellowish than brownish. The location of the USGS sensor (Baton Rouge public dock) is indicated by the (red) triangle, and the location of the Gybe sensor (The Water Institute of the Gulf) is indicated by the (blue) star.

(Zimmer et al., 2019), where various interventions (e.g., cover crops, bioreactors and floodplain restoration) are being implemented by NGOs, government and private actors to reduce nitrates (Piazza et al., 2015). Specifically, we sought to establish a link between optical reflectance (colour) data in the visible spectrum (≈400–700 nm) and nitrates in the Baton Rouge region of the MR (Figure 1a). To do this, we co-located an above-water hyperspectral sensor, manufactured by Gybe (http://gybe.eco), with an in-water nitrate sensor in the MR at Baton Rouge, Louisiana (USGS 07374000), 30° 26′ 14.36″ N, 91° 11′ 33.9″ W). This site also has a long-term water quality dataset collected by LSU (described below). We also used water quality data from a USGS station approximately 140 km north of Baton Rouge at Natchez, Mississippi (31° 33′ 37.59″ N, 91° 25′ 7.4″ W) to test the spatial extendability of our method. Nitrate measurements from USGS have been available from Baton Rouge since 2012, and those at Natchez only began operations in September 2022. The two sites, Baton Rouge and Natchez, share common source waters with little mixing or dilution from extraneous inputs. In a companion paper, we will examine this method for data collected in a similar manner in Tensas Bayou in the Atchafalaya River Basin near Bayou Sorrel, Louisiana (30° 09′ 58.1″ N, 91° 21′ 10.3″ W), where we investigate extending the modelling to heterogeneous waters consisting of distinct optical signatures and using the nitrate sensor at Morgan City, Louisiana (USGS 07381600). However, we first focused on the waters of the Lower MR to establish the feasibility of the method.

2.2 | Modelling approach

Our method to remotely estimate nitrate concentrations from remote sensing imagery uses a two-step procedure (sensu Stumpf et al., 2016). In Step 1, in-water nitrate measurements are correlated to common water quality parameters collected with the hyperspectral sensor (in situ data), and in Step 2, the in situ data are correlated to satellite estimates of the same water quality parameters. The overall result is an estimate of nitrate concentration from operational satellite water quality products such as chlorophyll-a and turbidity concentrations estimated from Sentinel-2 (Vanhellemont & Ruddick, 2016). The correlations are site (and region) specific, but they allow us to extrapolate from point nitrate data along a river to infer nitrate concentrations in the region of the point sampling.

The quality of the models depends on the quality, quantity and consistency of the input data. Therefore, we used three different datasets to investigate models that correlate visible water parameters and nitrates. The first dataset contains lab measurements from a long-term study by LSU (Turner et al., 2022). The second dataset contains observations from an above-water hyperspectral sensor, manufactured by Gybe (http://gybe.eco). This sensor was vicariously calibrated for water quality parameters (e.g., turbidity) with a co-located USGS gauge. The LSU data represent a deep historical record (23 years) and presumed high accuracy in constituent concentrations. In contrast, Gybe's hyperspectral data record is shorter (3 years at this location) and contains data only during daytime, because the sensor depends on sunlight. Both datasets include turbidity and chlorophyll-a concentrations. We combined the two datasets to make a single training dataset, where the co-located USGS turbidity gauge measurements were used to cross-calibrate the turbidity values between the LSU measurements and the Gybe sensor data. A cross-correlation for chlorophyll-a was not possible, because the USGS gauge does not have a chlorophyll-a sensor. Therefore, the chlorophyll-a values for these two datasets were combined without adjustment.

2.3 | In situ data

We installed the Gybe, above-water optical sensor in Baton Rouge during Fall 2019 to measure a suite of water quality parameters (e.g., turbidity and chlorophyll-a concentration) from the above-water remote sensing reflectance (Rrs). This sensor contains two spectrometers. One faces upwards, with a diffuser measuring the downwelling irradiance, and one faces the water and measures the water-leaving radiance. Both spectrometers are calibrated over a spectral range of approximately 400–700 nm with >200 spectral bands. The Rrs is estimated from the ratio of upwelling radiance to downwelling irradiance, with a sky glint correction using the 3c algorithm (Groetsch et al., 2017). The sensor is located on the dock of The Water Institute of the Gulf (Figure 2), approximately 300 m downstream from where a nitrate-plus-nitrite sensor has measured hourly data since 2012 (USGS 07374000). This is also the site where the LSU lab-analysed samples were collected for biogeochemical parameters between 1997 and 2018 (Turner et al., 2022). The LSU dataset consists of 866 values from all seasons, and the Gybe dataset reports 294 values. The Gybe sensor typically reports data at 15-min time intervals; however, in this study, we used only one summary measurement per day and only values that passed strict quality control for issues like glint or variable cloud cover are used. The combined dataset consists of 1160 values of turbidity, chlorophyll-a and nitrates plus nitrates between 2004–2018 (LSU) and 2020–2022 (Gybe). Additionally, because we were interested on the effect of river discharge on model fit, we included a dataset with contemporaneous USGS discharge values across this same time period.

2.4 | Satellite data

We obtained top of atmosphere radiances (Lla) from the Sentinel-2 satellite from the European Space Agencies (ESA) Copernicus Hub from January 2016 to January 2023. From 2016 until Fall 2017, this collection consists only of Sentinel-2A with a 10-day revisit time. Forward from that date the revisit time is 5 days, corresponding to when Sentinel-2B came online. Sentinel-2 imagery was processed to surface reflectances, using the open-source code Acolite (Vanhellemont & Ruddick, 2016). Satellite estimates of turbidity and chlorophyll-a concentration were computed using the algorithms of Nechad and Mishira, respectively (Mishra & Mishra, 2012; Nechad et al., 2010). The Nechad algorithms use a global water database for the satellite estimates.

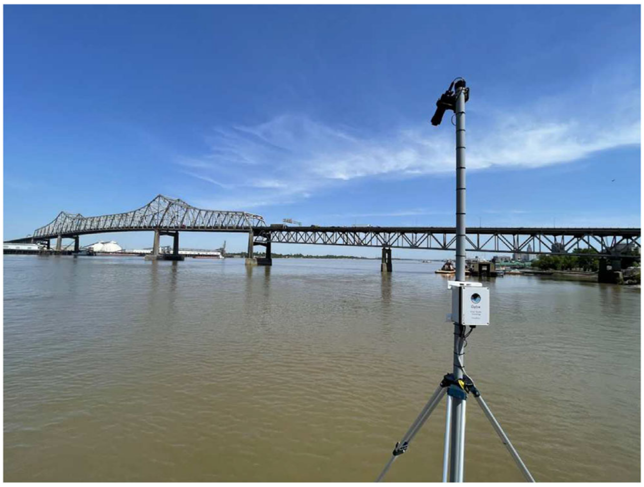

FIGURE 2 Picture of the Gybe Sensor looking over the Mississippi at The Water Institute of the Gulf with the Horace Wilkinson Bridge in the background.

In this study, we adjusted the parameter values using the in situ turbidity data available from USGS gauges to insure consistency between the Sentinel-2 products and the training dataset. This process is called vicarious calibration in remote sensing studies because it uses local data for the correlations instead of a global database and consists of a linear fit providing a gain and offset adjustment (Murakami et al., 2022). The full image set consisted of 411 images of which 91 were of sufficient quality (cloud and glint free) for use in this study.

2.5 | Surrogate modelling

Our method uses a surrogate modelling approach to estimate nitrate concentrations with remote sensing. Surrogate modelling is a term used by USGS to describe the estimation of a water quality parameter (e.g., phosphorus concentration) using correlation to other water parameters, like turbidity and discharge (Rasmussen et al., 2005). The utility of this approach is that the target parameter is usually more difficult or expensive to measure then the source parameters. For instance, turbidity gauges are both less expensive and more reliable than nitrate gauges. Specific surrogate models are empirical and site specific. The mathematical problem is the estimation of a target variable, y,

$$y = f(x_1, x_2, \ldots, x_n), \tag{1}$$

in terms of multiple source variables $x_1, \ldots, x_n$, given a discrete sample set for both input (x_n) and output variables (y). We call the above problem static because it has no feedback (autoregressive variables). However, in our approach and the examples described in this paper, we also assume that the models are continuous, meaning that the target model is a surface (or more generally a manifold Ziemann, 2015), which can be estimated by explicitly formulating the underlying assumptions in the model and performing a linear or nonlinear regression method to estimate the model parameters.

The USGS guidelines for surrogate modelling assume linear models for the model function $f(\cdot)$. A linear framework provides well-tested statistical and optimization procedures for model selection and estimation. However, in this study, we extended the model types to include nonlinear functions, specifically splines. Nonlinear models better represent the curvature in underlying data, but the statistical guidelines for model selection and optimization are more challenging.

Generally, nonlinear optimization requires two considerations—one for model fit and one for model size. The spline models used here achieve a parsimonious model by pruning an initial model fit by removing terms that have the least impact on the model fit.

2.6 | Data normalization and preprocessing

Before we could apply the surrogate modelling approach, some translation of units was necessary to make the datasets consistent. For example, the LSU dataset contained total suspended solids (TSS) concentrations, which we translated to turbidity with a linear regression against USGS turbidity gauge values (see Figure 3). Similarly, we converted the chlorophyll values in the Gybe data to values consistent with the LSU dataset also using a linear regression. The Gybe spectrometer estimates a normalized difference chlorophyll index (NDCI) using the 665- and 708-nm bands with model coefficients that were initially calibrated, in part, with data from the MR Delta (Mishra & Mishra, 2012). During the 3-year overlap (2016–2018), four daily chlorophyll matches were found between the Gybe and LSU datasets. Despite the limited number of matches, these are used to apply a vicarious adjustment of $\mathrm{Chl(LSU)} = 0.72 * \mathrm{Chl(Gybe)}$ to bring the two datasets into alignment.

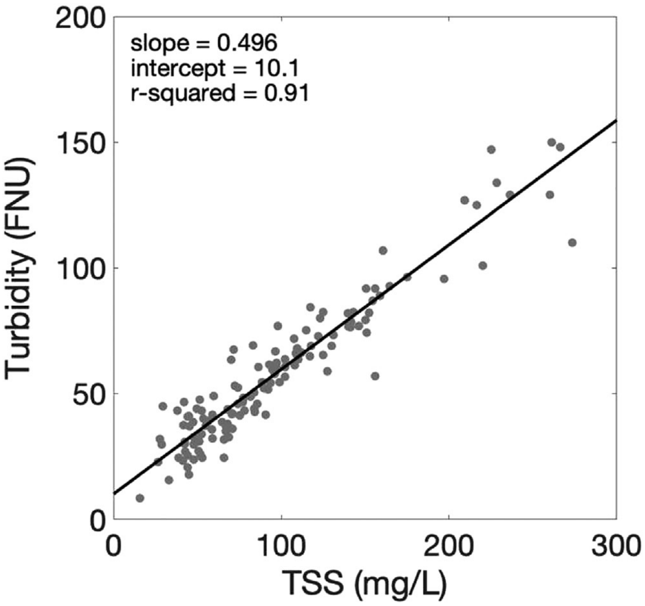

FIGURE 3 Calibration function mapping LSU measured total suspended solids (TSS, mg/L) to the USGS measured turbidity (FNU) at Baton Rouge, LA. There are 131 contemporaneous data samples between 2016 and 2018.

Nitrate values are also matched between the USGS and LSU datasets. The USGS reports nitrogen as nitrate plus nitrite ($\mathrm{NO_3 + NO_2}$), so we divided the LSU data by the molecular weight of nitrogen (14.01) to relate the equivalents to µmol. The LSU and USGS datasets showed excellent consistency as shown in Figure 4, and so, unlike the turbidity and chlorophyll values, no vicarious adjustment is added to bring these two datasets into alignment. We also added a seasonality variable to the training dataset captured by a sinusoidal variable:

$$s = (1 + \sin((2 * \pi * N) + \frac{3}{2}\pi))/2, \tag{2}$$

where $N = n/365$ is the normalized day of year.

The USGS time series dataset had overlapping periods between both the LSU and Gybe datasets and was used to build a consistent training dataset. Several preprocessing steps were performed when combining the datasets. First, the USGS dataset was broken into subsets whenever a data gap was greater than 3 h, and gaps less than 3 h were filled in by linear interpolation. Next, we smoothed the USGS data records to remove digitization noise. Data reported by USGS contain three significant digits. When raw time series data are plotted, there are small visible jumps in the data ('digitization noise'), which are artefacts of the effective digitization of the supplied data. To smooth the data, we used a moving mean filter with a window length of 12 h.

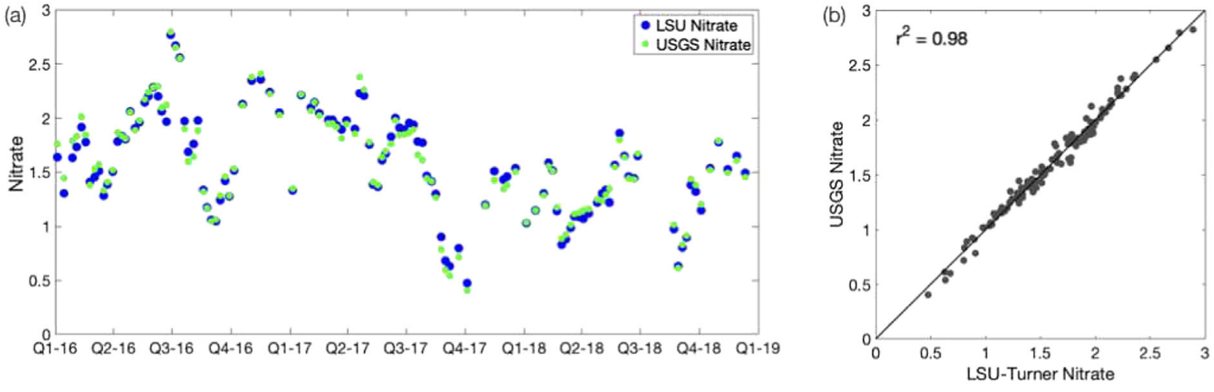

FIGURE 4 (a) Time series of LSU nitrate values measured from lab assays compared to an in situ nitrate gauge installed at Baton Rouge, LA (USGS 07374000). (b) The 1-1 plot comparing nitrate values showing that despite the two different measurements techniques, both datasets are consistent (with a correlation coefficient $r^2 = 0.98$) and can be combined to extend the nitrate time series.

Processed USGS turbidity and nitrate data were then matched to all LSU and Gybe data within a 1-h time window, and this training dataset was then used to estimate correlations between the input variables (turbidity, chlorophyll-a concentration, seasonality and possibly discharge) and the target variable, nitrate-plus-nitrite concentrations. As is common with hydrologic datasets, the training data were concentrated at the origin, so a Log (Base 10) transform was applied to the turbidity, chlorophyll-a, discharge and nitrate-plus-nitrite quantities before estimating correlations. Our final step before function fitting was to normalize the data to aid with numerical estimations. All the training time series data are inherently positive. So instead of centring on the mean, we normalized the data records to a magnitude of one with, $N(x) = x/\max(x)$.

2.7 | Model fitting

Using the normalized training dataset, we then created spline models to correlate nitrate concentration to the contemporaneous values for turbidity, chlorophyll-a concentration and seasonality. Specifically, we used multivariate adaptive regression splines (MARS; Friedman, 1991) as implemented in the Matlab ARESLab toolbox (Jekabsons, 2016). While we think there may be more robust function-fitting methods, such as neural nets, that we will explore in future work, we chose to use spline models, in part to aid with the presentation and interpretation of our surrogate modelling method. The spline model-fitting process consists of a forward operation, which finds knot locations and a backward pruning step that reduces model size, making final models as parsimonious as possible and easier to interpret. Additionally, we created correlation models with the addition of the discharge variable, because we wanted to compare model performance between models restricted only to satellite data and models augmented with commonly available USGS water quantity data. Typical training parameters started with cubic splines with a limit of 100 basis functions, which were reduced to approximately 50 after backward pruning.

All inputs to a model were evaluated with both training data and satellite-derived input products for turbidity and chlorophyll-a (Sentinel-2). As a simple metric for the quality of the fits, we examined variation about the one-to-one line between the target and predicted nitrate concentrations values (a simple linear regression) and reported an (r^2) correlation coefficient as an indicator of the relative quality of a model fit. Finally, the USGS dataset was also used to vicariously calibrate turbidity derived from Sentinel-2. We estimated that a linear multiplier of 0.96 brought them into better correspondence. Because there is no in-water chlorophyll sensor at the USGS gauge, we were not able to vicariously calibrate the Sentinel-2 derived chlorophyll-a product.

3 | RESULTS

To provide context for the surrogate modelling results, consider the example imagery and data plots in Figure 1. The centre image (Figure 1b) shows a clear Sentinel-2 imagery of Baton Rouge from

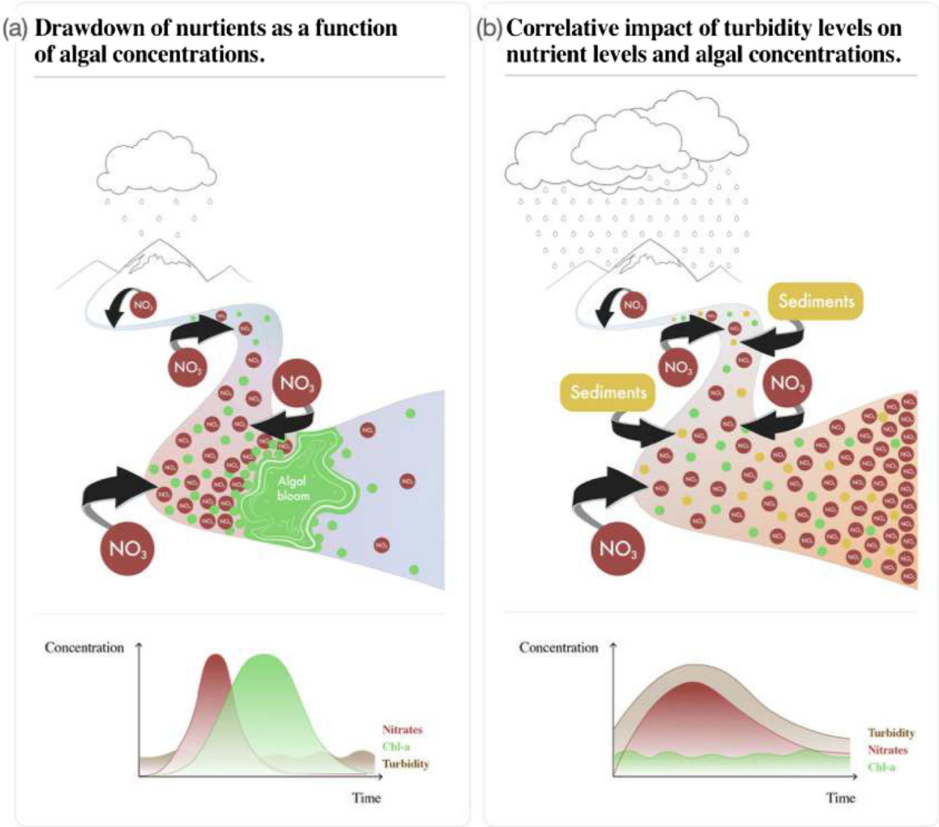
(a) Drawdown of nurtients as a function of algal concentrations.
NO3
Algal bloom
Concentration
Time
Nitrates
Chl-a
Turbidity
(b) Correlative impact of turbidity levels on nutrient levels and algal concentrations.
Sediments
NO3
Concentration
Time
Turbidity
Nitrates
Chl-a

FIGURE 5 A schematic of processes modulating nitrate concentrations in a river: (a) assimilation by phytoplankton (biological uptake) leading to a rapid decrease in nitrate concentrations and (b) land run-off increases sediment concentrations and hence turbidity, presumably proportional to nitrate concentrations (particularly in agricultural regions) leading to an increase in river nitrate concentrations.

18 October 2017. The deep green of the MR indicates a high chlorophyll content in the water. The right image (Figure 1c) is from 31 January 2018. Conversely, the brownish colour of the water indicates a large amount of suspended solids. The remote sensing imagery is used to translate these colour variations to estimates of target products (e.g., chlorophyll-a concentrations or turbidity).

As illustrated in Figure 5, our hypothesis is that nitrate concentrations can be modulated by both sediment and chlorophyll-a concentrations due to agricultural run-off and phytoplankton assimilation, respectively. Figure 5a illustrates how phytoplankton assimilation can decrease nitrate concentrations, while turbidity can increase nitrate concentrations (Figure 5b). Other recent work found these relationships to hold in the homogenous waters of the Kansas River where a data-driven model was demonstrated both these processes modulating nitrate concentrations. Specifically, it was shown that, relative to variations in sediment concentrations, the assimilation of nitrates is nonlinear (Tufillaro, 2023). Increases in chlorophyll (an indicator of phytoplankton abundance) led to a sharp decrease in nitrates, and the model was able to track the summer and fall boom-and-bust cycle of phytoplankton blooms, which caused oscillations in nitrate concentrations. Unlike the Kansas River, the Mississippi has very heterogenous source waters. However, we hypothesize that we could see a muted version of the modulation of nitrates by sediments and phytoplankton reflecting process in Mississippi itself or from more homogenous upstream source waters which get diluted by mixing with other source waters. Figure 6 shows a simple spline model of the LSU dataset, which appears to support this hypothesis. A response surface fitting the point cloud for input variables of turbidity and chlorophyll-a concentration, and an output of nitrate concentration, shows an increase in nitrate concentrations at low chlorophyll concentrations and rising turbidity, while nitrate concentrations appear to fall rapidly with increasing chlorophyll-a concentrations at lower turbidity.

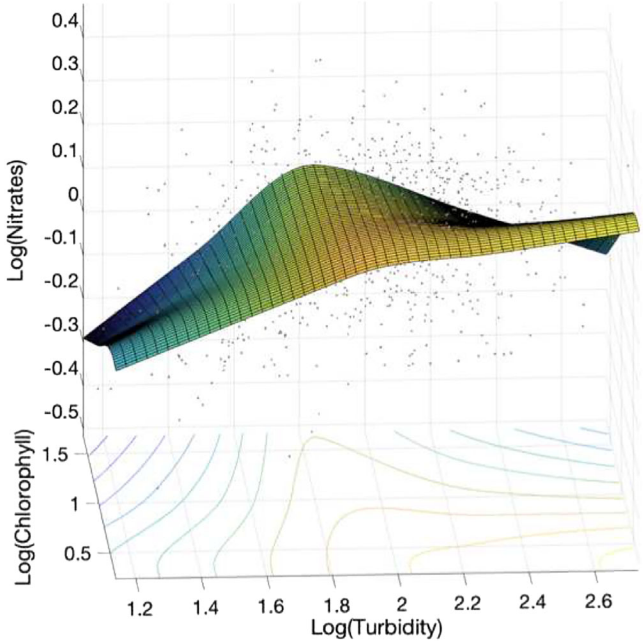

FIGURE 6 A spline regression of the point cloud from the LSU dataset with input variables $\log_{10}$ of turbidity and chlorophyll-a and output $\log_{10}$ of nitrate concentration.

As stated in the methods, we used multivariate splines to estimate the nonlinear response surface generated by the training set with possible input variables of turbidity, chlorophyll-a concentration, seasonality and discharge. Specifically, we started with linear splines, and the main tuning parameter we adjusted was the initial starting value for the number of spline knots which effectively controls the model size (values range from 1 to 100 or a maximum value of approximately 1/10 of the training set). After optimization, the model typically had about 50 parameters.

While we experimented with models using different sets of input variables and model parameters to estimate models that both satisfy low errors on the training set (as indicated by correlation coefficient r^2), the best model performance was achieved with the full training dataset and a complete set of input variables (seasonality, discharge, turbidity and chlorophyll-a concentration), which were nonlinearly regressed to the LSU/USGS nitrate-plus-nitrate concentration.

Figure 7 shows the one-to-one plots and correlation coefficients generated by two different training sets. The use of a discharge variable regressor improved the validation test set ($r^2 = 0.55$, Figure 7a; $r^2 = 0.65$, Figure 7b). Figure 7c indicates that the surrogate model developed using the terrestrial training set has utility for satellite-derived turbidities and chlorophyll-a inputs as well, as indicated by a correlation coefficient of $r^2 = 0.42$. Though this is a modest correlation, we expect that the result can be significantly improved as more satellite data imagery becomes available leading to improved vicarious calibration estimations between the satellite data products and the in situ measurements, as well as improvements to the atmospheric correction processing.

To test the hypothesis of the spatial extendibility of surrogate models for nitrates, we used the spline surrogate data model created for Baton Rouge, LA, to estimate nitrate values at the USGS nitrate gauge at Natchez, MS, about 140 km north of Baton Rouge. As a first step, we check for any correlations seen at these two sites based on the USGS in situ nitrate sensor data. Figure 8a shows the time series of nitrate concentration from September 2022 to January 2022. Note that the Natchez nitrate sensor only came online at the end of September 2022 and also has data gaps in November and small gaps (e.g., mid-January) later in 2022. The Baton Rouge nitrate sensor has a large gap also in September. Thus, the only overlap between all three datasets during 2022 is a short period at the end of September. Comparing sensor signals at these two sites, we see that the Baton Rouge sensor time series clearly lags the Natchez sensor time series (Figure 8a). We estimate this lag as ≈63 h by cross correlating the two time series. This translates to a mean current of 2.2 km/h assuming a 140-km distance. Historical averages for the MR velocity range between 1 and 5 km/h from the headwaters to the mouth near New Orleans, Louisiana. After adjusting the time series by 63 h, the nitrate values at both sites are in good one-to-one correspondence (Figure 8b), with the nitrate concentration at Baton Rouge showing

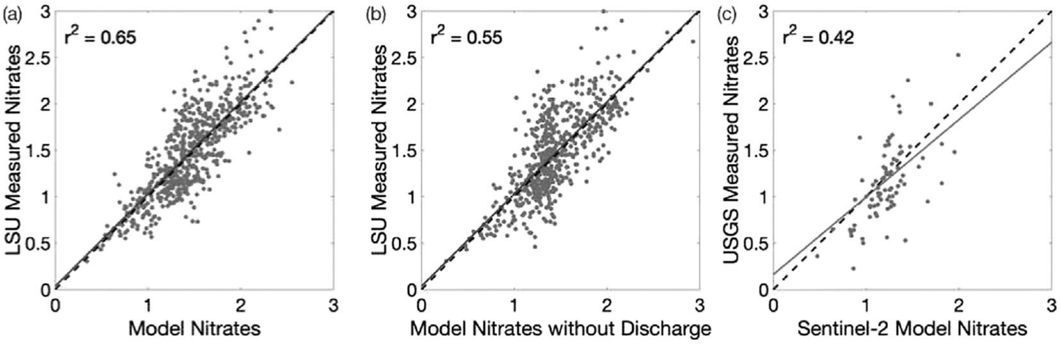

FIGURE 7 One-to-one maps for measured values of $NO_2 + NO_3$ and the surrogate model predictions. The relative quality of the model is indicated by the correlation coefficient r^2 with $r^2 = 1$ indicating a perfect correlation and $r^2 = 0$ indicating a correlation no better than an estimate of the mean value of the training set. (a) Surrogate model results for a *validation training* data using all the input regressors: seasonality, discharge and turbidity and chlorophyll-a, and the target variable is nitrates plus nitrates. (b) Surrogate model results for a *validation training* set using only input regressors for seasonality, turbidity and chlorophyll-a. (c) Surrogate model results for a *test set* using inputs as seasonality, discharge and Sentinel-2 (satellite-derived) product values for turbidity and chlorophyll-a concentrations.

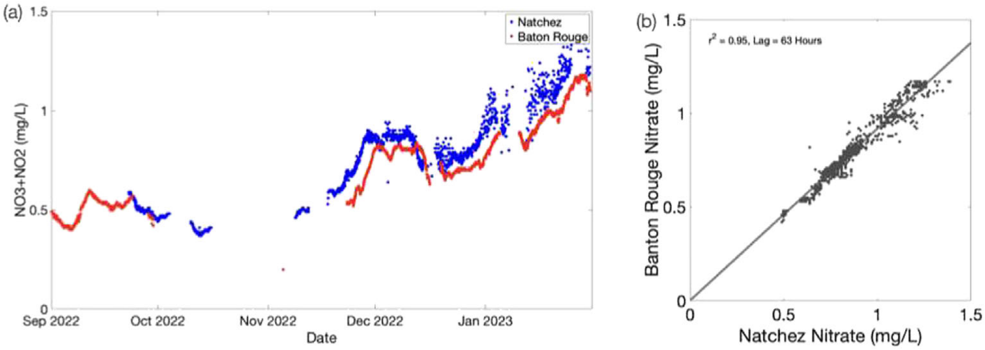

FIGURE 8 (a) Time series of $NO_2 + NO_3$ at Natchez, MS, and Baton Rouge, LA (no time delay). (b) The one-to-one map of the time series at Natchez, MS (delayed 63 h), and Baton Rouge, LA, showing a linear relation with a 9% decrease in concentration downstream between the two USGS in situ nitrate sensors.

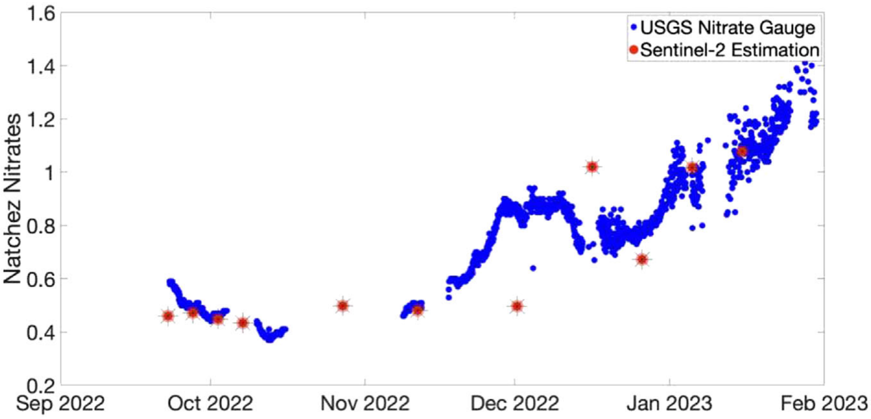

FIGURE 9 Time series of $NO_2 + NO_3$ at Natchez, MS, compared to values predicted from remote sensing using a regional model trained on data from Baton Rouge, LA.

approximately 9% lower nitrate values than at Natchez. The cause of the difference (such as dilution, nutrient consumption or differences in calibration constants) has yet to be determined.

The results of applying Baton Rouge surrogate data model to nitrate values derived from Sentinel-2 imagery at Natchez, MS, is shown in Figure 9. From October to January 2022, there are 11 satellite retrievals (clear satellite images over Natchez, MS), and though there is not enough satellite matches to date to make a statistical evaluation, the data trend does show an apparent correlation between the model predictions and the rise in nitrate values measured by the USGS gauge. We note that there are underestimations and overestimations at dates in the first half of December which might be explained by a turbidity bump that is not well correlated with data in the training set.

4 | DISCUSSION

We successfully applied a surrogate modelling approach that used remote sensing data to estimate nitrogen concentration in the well-mixed waters of the Lower MR. We also found that we were able to use our model and satellite-derived optical data parameters to successfully estimate nitrogen concentrations 140 km upstream from the original gauging site. While there is much room for improvement in the model, our results agreed with earlier work in the Kansas River (Tufillaro, 2023).

There are two types of predictions implicit in this modelling approach, temporal and spatial. As discussed above, temporal prediction is best handled with a state space (dynamic) model, which is not practical in this instance because of the sparse and irregular sampling available from remote sensing data. The fact that a limited degree of temporal prediction is possible, based on historical data and static modelling, is a bit surprising. The temporal prediction is presumably based, in part, on repetitive seasonal (e.g., storms) and land use (e.g., fertilizer application in the spring) patterns, though this is just a hypothesis at present, but there is evidence for this hypothesis in the homogenous source waters contributing to the Mississippi (Tufillaro, 2023). However, a spatial model prediction of water parameters at an associated temporal instance (i.e., adjusted for Lagrangian transport) is not unexpected, particularly when there are minimal new inflows along the reach. Indeed, we expect that the spatial correlation would be even better if the model was able to utilize contemporaneous hyperspectral data from the late Fall and Winter of 2022 at Baton Rouge, which is missing only due to an unexpected interruption in the in situ sensor operations.

The importance of a suitable choice of regression variables is also illustrated in this study. The higher correlation coefficient (r^2) for models with discharge (Figure 4) we hypothesize is due to the fact that the discharge (and seasonality) variables help spread out the input space so that the variance in the mapping to the output space is lower. Mathematically, the map from the input space to the output space is subject to ambiguity due to overlapping data clusters; however (assuming some deterministic correlations between the variables), it is possible to reduce the output variance by different combinations of input variables and transformations of those variables. This more geometric perspective on modelling generally goes under the rubric of manifold learning (Bachmann et al., 2005).

In summary, this study advances previous work on surrogate modelling of nitrates in river in two directions. First, we illustrate how the use of nonlinear fitting functions can extend the toolkit for modelling correlations between key environmental parameters, such as discharge, turbidity and nitrates. Second, we are suggesting that the use of remote sensing and surrogate data correlations provides a tool for gauging the spatial variations of nutrients, and other biogeochemical parameters, that are not directly accessible by visible spectroscopy. This second conclusion is not a large leap from the current use of surrogate methods that correlate lab samples to in situ sensors. Indeed, a primary reason for the development of ocean colour satellites and algorithms is specially to gauge global carbon up take by ocean processes. The primary target variable is not chlorophyll-a concentration in that case but primary production. What we are suggesting here is using remote sensing for similar applications on a finer spatial scale.

This study also demonstrates that monitoring of outcomes is increasingly possible in conservation (Rissman & Smail, 2015). Conservation organizations, including nongovernmental and governmental organizations, most often report on outputs (e.g., acres conserved and miles of buffers installed) rather than outcomes (e.g., land cover and water quality) (Rissman & Smail, 2015). The Government Performance and Results Act of 1993 (updated 2010) requires federal agencies to report results (Frederickson & Frederickson, 2006, as cited in Rissman & Smail, 2015). Yet, agencies face data limitations and technical challenges when it comes to reporting outcomes. While these results are promising, we also acknowledge that multiple data sources and approaches will likely be important for assessing progress towards policy goals.

ACKNOWLEDGEMENTS

We thank Prof. R. Eugene Turner for his generosity in taking the time to explain many details regarding the LSU dataset. Funding for a portion of this study was provided by Enterprise Rent-A-Car Foundation.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in USGS National Water Information System at https://waterdata.usgs.gov/nwis/rt/. These data were derived from the following resources available in the public domain: USGS Gauge, https://waterdata.usgs.gov/monitoring-location/07374000/#parameterCode=00065&period=P7D.

ORCID

Nicholas Tufillaro https://orcid.org/0009-0006-2896-8832

REFERENCES

Bachmann, C. M., Ainsworth, T. L., & Fusina, R. A. (2005). Exploiting manifold geometry in hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, *43*(3), 441–454.

Bennett, M. G., Fritz, K. A., Hayden-Lesmeister, A., Kozak, J. P., & Nickolotsky, A. (2016). An estimate of basin-wide denitrification based on floodplain inundation in the Atchafalaya River Basin, Louisiana. *River Research and Applications*, *32*, 429–440.

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.

Diaz, R. J., & Rosenberg, R. (2008). Spreading dead zones and consequences for marine ecosystems. *Science*, *321*(5891), 926–929. https://doi.org/10.1126/science.1156401

Edwards, A. C., Hooda, P. S., & Cook, Y. (2001). Determination of nitrate in water containing dissolved organic carbon by ultraviolet spectroscopy. *International Journal of Environmental Analytic Chemistry*, *80*, 49–59.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, *19*(1), 1–67.

Greenhalgh, S., & Sauer, A. (2003). Awakening the dead zone: An investment for agriculture, water quality and climate change.

Groetsch, P. M., Gege, P., Simis, S. G., Eleveld, M. A., & Peters, S. W. (2017). Validation of a spectral correction procedure for sun and sky reflections in above-water reflectance measurements. *Optics Express*, *25*(16), A742–A761.

Jekabsons, G. (2016). ARESLab: Adaptive regression splines toolbox for Matlab/Octave, 2011. http://www.cs.rtu.lv/jekabsons

Johnston, C. A. (1991). Sediment and nutrient retention by freshwater wetlands: Effects on surface water quality. *Critical Reviews in Environmental Control*, *21*, 491–565.

McLellan, E., Robertson, D., Schilling, K., Tomer, M., Kostel, J., Smith, D., & King, K. (2015). Reducing nitrogen export from the Corn Belt to the Gulf of Mexico: Agricultural strategies for remediating hypoxia. *Journal of the American Water Resources Association*, *51*, 263–289.

Mishra, S., & Mishra, D. R. (2012). Normalized difference chlorophyll index: A novel model for remote estimation of chlorophyll-a concentration in turbid productive waters. *Remote Sensing of Environment*, *117*, 394–406.

Murakami, H., Antoine, D., Vellucci, V., & Frouin, R. (2022). System vicarious calibration of GCOM-C/SGLI visible and near-infrared channels. *Journal of Oceanography*, *78*(4), 245–261.

Nechad, B., Ruddick, K. G., & Park, Y. (2010). Calibration and validation of a generic multisensor algorithm for mapping of total suspended matter in turbid waters. *Remote Sensing of Environment*, *114*(4), 854–866.

Piazza, B. P. (2014). *The Atchafalaya River Basin: History and ecology of an American wetland*: Texas A&M University Press.

Piazza, B. P., Allen, Y. C., Martin, R., Bergan, J. F., King, K., & Jacob, R. (2015). Floodplain conservation in the Mississippi River Valley: Combining spatial analysis, landowner outreach, and market assessment to enhance land protection for the Atchafalaya River Basin, Louisiana, USA. *Restoration Ecology*, *23*, 65–74.

Rabalais, N. N. (2011). Troubled waters of the Gulf of Mexico. *Oceanography*, *24*, 200–211.

Rabalais, N. N., Turner, R. E., & Wiseman, W. J. (2002). Gulf of Mexico hypoxia, a.k.a. "the dead zone". *Annual Review of Ecology and Systematics*, *33*, 235–263.

Rabotyagov, S. S., Campbell, T. D., White, M., Arnold, J. G., Atwood, J., Norfleet, M. N., Kling, C. L., Gassman, P. W., Valcu, A., Richardson, J., Turner, R. E., & Rabalais, N. N. (2014). Cost-effective targeting of conservation investments to reduce the northern Gulf of Mexico hypoxic zone. *Proceedings of the National Academy of Sciences*, *111*, 18530–18535.

Rabotyagov, S. S., Kling, C. L., Gassman, P. W., Rabalais, N. N., & Turner, R. E. (2014). The economics of dead zones: Causes, impacts, policy challenges, and a model of the Gulf of Mexico hypoxic zone. *Review of Environmental Economics and Policy*, *8*, 58–79.

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. (No. 2005-5165).

Rissman, A. R., & Smail, R. (2015). Accounting for results: How conservation organizations report performance information. *Environmental Management*, *55*, 916–929.

Smith, M. D., Oglend, A., Kirkpatrick, A. J., Asche, F., Bennear, L. S., Craig, J. K., & Nanceet, J. M. (2017). Seafood prices reveal impacts of a major ecological disturbance. *Proceedings of the National Academy of Sciences*, *114*, 1512–1517.

Stumpf, R. P., Davis, T. W., Wynne, T. T., Graham, J. L., Loftin, K. A., Johengen, T. H., Gossiaux, D., Palladino, D., & Burtner, A. (2016). Challenges for mapping cyanotoxin patterns from remote sensing of cyanobacteria. *Harmful Algae*, *54*, 160–173.

Tufillaro, N. (2023). A manifold learning perspective on surrogate modeling of nitrates in the Kansas river. *Water Practice and Technology*. Manuscript submitted for publication.

Turner, R. E., Milan, C. S., Swenson, E. M., & Lee, J. M. (2022). Peak chlorophyll a concentrations in the lower Mississippi River from 1997 to 2018. *Limnology, and Oceanography*, *67*(3), 703–712.

Vanhellemont, Q., & Ruddick, K. (2016). Acolite for Sentinel-2: Aquatic applications of MSI imagery. In *Proceedings of the 2016 ESA Living Planet Symposium, Prague, Czech Republic*, pp. 9–13.

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 (No. 2021-1018). US Geological Survey.

Ziemann, A. K. (2015). *A manifold learning approach to target detection in high-resolution hyperspectral imagery*: Rochester Institute of Technology.

Zimmer, M. A., Pellerin, B., Burns, D. A., & Petrochenkov, G. (2019). Temporal variability in nitrate-discharge relationships in large rivers as revealed by high-frequency data. *Water Resources Research*, *55*(2), 973–989.

How to cite this article: 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