RESEARCH ARTICLE

Understanding links between water-quality variables and nitrate concentration in freshwater streams using high frequency sensor data

Claire Kermorvant[1]*, Benoit Liquet[1,2,3], Guy Litt[4], Kerrie Mengersen[3,5], Erin E. Peterson[3,6], Rob J. Hyndman[3,7], Jeremy B. Jones Jr.[8], Catherine Leigh[3,9]

1 Le CNRS et l'Université de Pau et des Pays de l'Adour, Laboratoire de Mathématiques et de leurs Applications de Pau, Anglet, France, 2 School of Mathematical and Physical Sciences, Macquarie University, Sydney, New South Wales, Australia, 3 ARC Centre of Excellence for Mathematics and Statistical Frontiers, Brisbane, Queensland, Australia, 4 Battelle, National Ecological Observatory Network, Boulder, Colorado, United States of America, 5 School of Mathematical Sciences, Queensland University of Technology, Brisbane, Queensland, Australia, 6 Peterson Consulting, Brisbane, Queensland, Australia, 7 Department of Econometrics and Business Statistics, Monash University, Clayton, Victoria, Australia, 8 Institute of Arctic Biology and Department of Biology and Wildlife, University of Alaska Fairbanks, Fairbanks, Alaska, United States of America, 9 Biosciences and Food Technology Discipline and School of Science, RMIT University, Bundoora, Victoria, Australia

* claire.kermorvant@univ-pau.fr

Abstract

Real-time monitoring using *in-situ* sensors is becoming a common approach for measuring water-quality within watersheds. High-frequency measurements produce big datasets that present opportunities to conduct new analyses for improved understanding of water-quality dynamics and more effective management of rivers and streams. Of primary importance is enhancing knowledge of the relationships between nitrate, one of the most reactive forms of inorganic nitrogen in the aquatic environment, and other water-quality variables. We analysed high-frequency water-quality data from *in-situ* sensors deployed in three sites from different watersheds and climate zones within the National Ecological Observatory Network, USA. We used generalised additive mixed models to explain the nonlinear relationships at each site between nitrate concentration and conductivity, turbidity, dissolved oxygen, water temperature, and elevation. Temporal auto-correlation was modelled with an auto-regressive–moving-average (ARIMA) model and we examined the relative importance of the explanatory variables. Total deviance explained by the models was high for all sites (99%). Although variable importance and the smooth regression parameters differed among sites, the models explaining the most variation in nitrate contained the same explanatory variables. This study demonstrates that building a model for nitrate using the same set of explanatory water-quality variables is achievable, even for sites with vastly different environmental and climatic characteristics. Applying such models will assist managers to select cost-effective water-quality variables to monitor when the goals are to gain a spatial and temporal in-depth understanding of nitrate dynamics and adapt management plans accordingly.

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

Introduction

Nitrate is one of the most reactive forms of inorganic nitrogen in the aquatic environment [1] and an essential component of the nitrogen cycle supporting life on Earth. In rivers, sources of nitrate include atmospheric deposition, groundwater, surface runoff and the biological degradation of organic matter present in freshwater ecosystems. In addition, anthropogenic sources associated with agricultural, industrial and urban land use are becoming increasingly prevalent [2]. This includes the combustion of fossil fuels, which contributes substantially to atmospheric deposition [3]. In its bio-available form, nitrate is assimilated for growth and metabolism by riverine biota (e.g. algae, macrophytes and some bacteria) that form the basal components of aquatic food webs [1]. However, an excess of nitrate can lead to problems associated with eutrophication, such as decrease in light infiltration and dissolved oxygen concentration [4, 5]. This can negatively impact the health of aquatic biota such as invertebrates and fish [6–8]. Understanding the dynamics of nitrate concentration in rivers, and the relationships nitrate has with other water-quality variables, is therefore of primary importance for the effective management of freshwater ecosystems.

Monitoring is central to understanding the links between water-quality variables and the health of freshwater ecosystems [9]. Advances in the development of *in-situ* environmental sensors have led to their world-wide and long-term use in environmental monitoring [10–12]. Yet, water-quality monitoring in rivers still relies primarily on the time-consuming and expensive manual collection of samples, with the resultant data being sparse in space and time [13]. Fortunately, relatively low-cost *in-situ* sensors and sampling methods are being developed that allow some properties such as water temperature, turbidity, oxygen concentration, salinity and conductivity to be semi-continuously sampled and then analysed statistically [14, 15]. These high-frequency data-sets provide unique opportunities to better understand water-quality dynamics.

The large data-sets generated by *in-situ* sensors also present new challenges when analysing, modelling and reporting water-quality data [14–16]; for example in terms of quality assurance and control [13]. The prohibitive cost of certain sensors [17], such as optical sensors used to estimate high-frequency nitrate (NO_3^-) concentration, mean that they can only be deployed at a small number of sites and/or for limited periods of time. An additional challenge is to develop transferable models of water-quality dynamics and the links among water-quality variables for disparate river systems, especially for properties of interest like nitrate [18], especially when sensors are sparsely deployed. Ubiquity in water-quality relationships across climate zones and watersheds have thus far remained difficult to detect and model due to the many different processes that can be responsible for the resultant dynamics [19].

While several other studies have explored models for nitrate concentration in rivers, they have been developed for different purposes, many for prediction rather than to understand underlying relationships among water-quality variables, and with varying degrees of success. For example, [20] used Artificial Neural Networks to predict monthly values of nitrate from multiple measures, including concentrations of other nutrients measured using traditional sampling and laboratory analyses. More recently, [18] predicted nitrate from non-nutrient-based water-quality data collected from high-frequency, *in-situ* sensors using generalised-linear mixed-effects models (GLMMs) with a continuous first-order auto-regressive correlation (AR(1)) structure to account for temporal auto-correlation. However, GLMMs detect linear relationships and, as pointed out by [21], the relationships between nitrate and other water-quality variables tend to be nonlinear. [21] also investigated relationships between nitrate and other water-quality variables using high-frequency sensor data with the aim of prediction using Random Forests Regression (RFR) models, which can handle nonlinear interactions

among variables. However, predictions could not be extrapolated beyond the ranges of the input data and the resultant structure of the models was relatively opaque [21]. Links between discharge (also known as flow) and nitrate have also been investigated. However this relationship is complex, not least due to nutrient spiralling in streams [22], with some studies showing poor ability to model and explain variation in nitrate [23, 24] while others finding much stronger relationships [25].

Our goal is to explore opportunities and address some challenges associated with high-frequency *in-situ* monitoring data. More specifically, we identified key variables and developed an additive model structure, which we used to understand the complex relationships between nitrate concentration and other water-quality variables (rather than as an exercise in prediction) collected in disparate climatic regions and subject to different levels of anthropogenic impacts.

Materials and methods

NEON database and water-quality sensors

The National Ecological Observatory Network (NEON) database provides open data from sites across the United States of America (USA). All NEON sites are equipped with high-frequency sensors and follow standardised configuration, calibration and preventive maintenance procedures [26, 27], with *in-situ* measurements and sample analyses following protocols [28]. Nitrate is measured in μM using a 10 mm path length SUNA V2 UV light spectrum sensor. The SUNA V2 collects data reported as a mean value from 20 measurements made during a sampling burst every 15 minutes (Table 1). Other sensors collect specific conductance (μS/cm), dissolved oxygen (mg/L), temperature (˚C) and turbidity (Formazin Nephelometric Units, FNU) data as one-minute instantaneous measurements. Surface water elevation (i.e. meters above sea level) data are also recorded as five-minute instantaneous measurements (Table 1).

Study sites and time-series data

We extracted time-series of nitrate [29] and other surface water-quality variables [30–32] from three different sites within the NEON database on 29 January 2021 (see Table 1): the Arikaree River in Colorado, Caribou-Poker Creeks Research Watershed in Alaska, and Lewis Run in Virginia, which we will refer to as Arikaree, Caribou and Lewis Run, respectively throughout (Table 2). As nitrate measurements were collected less frequently than other water-quality variables (see Table 1), all sensors took measurements each time nitrate was sampled. Water-

Table 1. Details on NEON sensors, variables collected, units of measurement, associated data-collection intervals, and the NEON data product number for data used in this study.

Sensor	Water-quality variable	Unit	Raw sensor collection interval	Published interval	NEON data product number
SUNA v2	nitrate	μmol/L NO3-N	average of 20 bursts every 15 minutes	15 minute	DP1.20033.001
Level TROLL 500	water level	meters above sea level	1 minute	5 minute	DP1.20016.001
YSI EXO Optical Dissolved Oxygen	dissolved oxygen	mg/L DO	1 minute	1 minute	DP1.20288.001
YSI EXO Turbidity	turbidity	FNU	1 minute	1 minute	DP1.20288.001
YSI EXO Conductivity and Temperature	specific conductance	μ S/cm	1 minute	1 minute	DP1.20288.001
Platinum Resistance Thermometer	temperature	˚C	1 minute	1 minute	DP1.20053.001

https://doi.org/10.1371/journal.pone.0287640.t001

Table 2. Details of NEON sites used in this study.

	Arikaree	Caribou	Lewis Run
Watershed (km^2)	2,875	106	11.9
Latitude/Longitude (WGS 84)	39.758206, -102.44715	65.153224, -147.50397	39.095637, -77.983216
Manager	The Nature Conservancy	Bonanza Creek Long-Term Ecological Research Program and University of Alaska Fairbanks	Casey Trees (nonprofit organisation)
Climate zone	Semi-arid	Subarctic	Temperate
Land use / land cover	Grasslands, agriculture	Subarctic taiga, discontinuous permafrost	Fields, pastures, woodlands and small ponds
Flow persistence	Intermittent (dry in summer)	Perennial by ice covered in winter	Perennial
Mean annual precipitation (mm)	450	262	976
Urbanisation present in watershed?	No	No	Yes
Period of data analysed	January 2018 to December 2019	January 2018 to December 2019	January 2018 to December 2019
Period of data modelled	September 2018 to December 2019	June 2018 to October 2019	January 2018 to December 2019

https://doi.org/10.1371/journal.pone.0287640.t002

quality measurements from other time points were therefore discarded so that we only used data measured every 15 minutes (i.e. at each nitrate-measurement time point) in the analyses. Prior to analyses, we also removed data labelled as anomalous (e.g. due to known sensor calibration problems) during the rigorous NEON quality assurance and quality control procedure [16]. This, along with periods of missing data resulting from flow intermittence or sensors being temporarily out of service, meant that the time series of data from each site differed. For example, the river at Caribou freezes from approximately October to May each year and NEON removes most sensors from the site to prevent damage or loss. Despite these gaps in the data, at least 50% of the time series we examined from any one site overlapped with that of the other sites. Table 2 provides details about the selected data time series for each site, along with the actual time-period that could be used for modelling due to missing data. The time series of data are also available in the S1 File.

Statistical analyses

NEON publishes many environmental data products. The "water-quality" data product [30] includes high-frequency pH, dissolved oxygen, oxygen saturation, turbidity, specific conductance, conductivity, chlorophyll-*a*, and fluorescent dissolved organic matter (fDOM) data streams. The "temperature (PRT) in surface water" product [31] contains high-accuracy temperature data and the "surface water elevation" product [32] includes elevation data derived from pressure transducers and site-surveyed elevations. Among the water-quality variables, many exhibit strong correlation with each other. We investigated multicollinearity in order to select only those variables that were independent or weakly correlated with each other. Multicollinearity between covariates can influence parameter estimates and inflate variances, leading to improper inference from fitted models [33], especially when the sample size is small. Although the data sets in this study were large (n = 70080 at Caribou, Arikaree and Lewis Run), we checked for multicollinearity using the variance inflation factor (VIF) [34] to identify and remove any covariates that were strongly multi-collinear (VIF < 6). Based on this rule, we decided to not use conductivity, fluorescent dissolved organic matter (fDOM) pH and oxygen saturation. Also, the chlorophyll-*a* data were not taken into account in this study due to there

being an excessive number of anomalies. Therefore, we considered the following variables for further study: elevation of surface water, temperature, specific conductance, dissolved oxygen and turbidity.

For the purposes of statistical analyses, we considered nitrate as the response (i.e. dependent variable) and the other water-quality variables (see Table 1) as covariates (i.e. explanatory variables or predictors). We included a continuous covariate of time to account for the natural temporal variability that can occur in water quality variables, like nitrate concentration, in rivers. Visual examination of the distributions of the response and covariates indicated that turbidity had a strongly right-skewed distribution and was therefore log-transformed (i.e. log (turbidity + 1)) prior to analysis.

Generalised additive mixed models (GAMMs) [35] were built to link nitrate concentration with covariates from each site individually as described by the equation:

$$Y_i = \beta_0 + \sum_{k=1}^{m} s_k(z_{ki}) + \eta_i \qquad (1)$$

where z_{ki} are covariates measured at the ith sample ($i = 1, \ldots, n$). Here, β_0 is an intercept and η_i is the auto-regressive integrated moving-average (ARIMA) (ρ, ϕ) error, $\eta_i = \varepsilon_i + \sum_{j=1}^{p} \rho_j \eta_{i-j} + \sum_{l=1}^{q} \phi_l \varepsilon_{i-l}$ with ρ_j the autocorrelation parameters and φ_l the moving average parameters, and $\varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$ is Gaussian white noise. The associated smooth function $s_k(\cdot)$ of each water-quality variable z_k was defined using thin plate spline regression [36].

This model defined in (1) is estimated using a two-step modelling framework:

1. Step one: A generalised additive model (GAM) is used to model potential non-linear links between nitrate concentration and covariates. A stepwise variable-selection procedure was implemented and the 'best' GAM (variables and penalisation of smooth splines) for each site was identified based on the Akaike Information Criterion (AIC) [37]. To avoid over-fitting of the GAM model, the maximal value of degrees of freedom of the smooth terms was fixed at 6. This value was chosen to be large enough to allow for shapes that could be explained by periodic ecological processes potentially occurring at intra-annual to annual time scales. There is an identifiability issue between time trends due to seasonal patterns, and autocorrelation due to time-changing covariates. The value we chose ensured the time trends were flexible enough to handle longer-term trends, while not so large as to overfit temporal variability and capture shorter-term fluctuations.

2. Step two: An autoregressive–moving-average model was fitted to the GAM residuals to take into account the structural dependence of the time series data (GAMM). The best ARIMA regression on the GAM residuals were identified, based on the Akaike Information Criterion, to account for temporal autocorrelation in the time series data.

We then assessed the statistical significance and importance of each covariate in the best models to better understand their effects on nitrate. Variable importance can be estimated easily with linear models using the partial R^2 (i.e. the proportion of variation explained by a covariate in a model), but this approach cannot be used in a GAMM [38]. Therefore, we compared the deviance explained by the best GAMM and the same GAMM with each water-quality covariate iteratively removed. We choose to present deviance explained in this paper to compare models and variable importance at each site because deviance explained is a percentage (restricting it between 0 and 100) making it easy to interpret. Finally, to compare performance of GAM versus GAMM in each site, we assessed the approximate Akaike Information

Criterion (aAIC) given by the formula:

$$aAIC = n \log(\hat{\sigma}^2) + 2k \qquad (2)$$

[39] where n is the length of the time series, $\hat{\sigma}^2$ is the variance of the model residuals and k is the total number of degrees of freedom in the model. We chose this approach because it is computationally efficient compared to other cross-validation methods (e.g. with multiple training/test set splits) and has reliable convergence properties asymptotically equivalent to one-step time series cross-validation. The smaller the aAIC, the better the model performance.

All analyses were undertaken in R statistical software using the `car` [40], `gam` [41], `mgcv` [36], and `forecast` [42] packages. The R script used to implement the analyses is provided in the GitHub repository: https://github.com/Claire-K/nitrate_links.

Results

Water-quality characteristics within and among sites

Each site had distinct water-quality characteristics (Figs 1 and 2). Lewis Run had a higher nitrate concentration than Caribou and Arikaree sites (median = 5.5, 28.4 and 192 μM at Arikaree, Caribou and Lewis Run, respectively). For specific conductance, the Caribou site differed from the two other sites having, lower values (median = 531.6, 77.15 and 577.8 μS/cm at Arikaree, Caribou and Lewis Run, respectively). Dissolved oxygen concentration also differed among the three sites (median = 7.3, 12.34 and 9.55 mg/L at Arikaree, Caribou and Lewis Run, respectively). Water-quality variables exhibited more variability at Arikaree and Lewis Run than at Caribou (Fig 1). Temperature ranges were narrowest at Caribou (0 to 13˚C), wider at at Lewis Run (1 to 22˚C), and widest Arikaree (0 to 34˚C). As noted above, turbidity was strongly right-skewed in distribution at all three sites, with the mean always greater in value than the third quartile (mean = 95.45 FNU, Q3 = 9.98 FNU at Arikaree; mean = 3.41 FNU, Q3 = 3.15 FNU at Caribou; mean = 23.85 FNU, Q3 = 23.93 FNU at Lewis Run). Finally, surface water elevation was very different among sites, and despite the small ranges between minimum and maximum elevations, exhibited some temporal variability at Caribou (230.0–230.8 m), Arikaree (1179–1180 m) and Lewis Run (125.9–126.7) during the study period.

Diel fluctuations in water-quality occurred at all three sites despite differences in the distributions of water-quality data among them (top plots of Fig 2). At Caribou, nitrate, dissolved oxygen, specific conductance and turbidity increased while temperature decreased during the afternoon and the night. At Arikaree, the diel patterns exhibited comparatively more variation than at Caribou, and both nitrate and dissolved oxygen fluctuated in the opposite direction (i.e. decreasing at night and increasing during the day). Lewis Run exhibited similar diel patterns in turbidity as Caribou and Arikaree in terms of there being a clear alternation between day and night.

When a flow event (lower plots in Fig 2) occurred at Arikaree site (i.e. when the elevation level suddenly rose), nitrate concentration, oxygen concentration and turbidity increased, while specific conductance and temperature decreased. Conversely, at Lewis Run, a sudden increase in water level coincided with a decrease in nitrate concentration. At Caribou, the rise in water level was accompanied by an increase in turbidity, but the relationships between water level and the other water-quality covariates were more complex.

GAMs

GAM regressions were used to understand the links between nitrate concentration and each covariate (Fig 3). The smooth regressions between nitrate and the covariates revealed
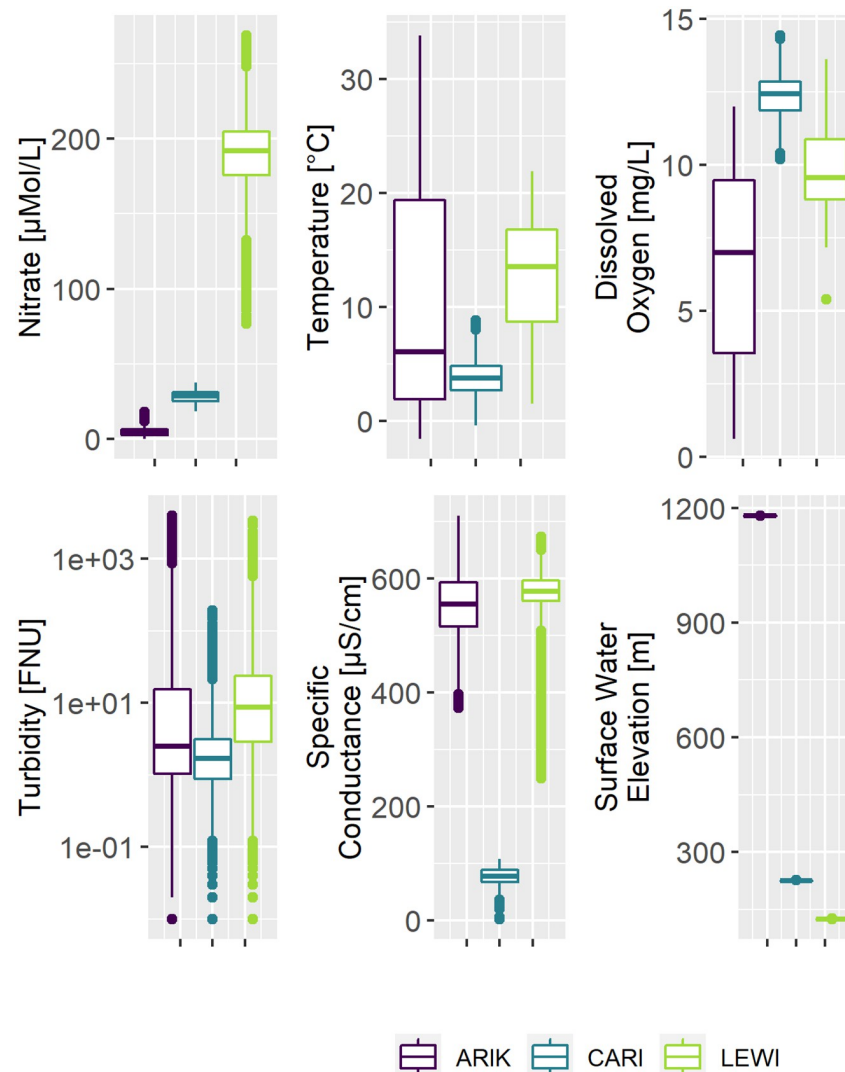
Fig 1. Box-plots of water-quality data for Arikaree, Caribou and Lewis Run. Bold lines within boxes represent medians and lower and upper edges of boxes represent the interquartile range (IQR), with whiskers extending to 1.5 times the IQR. Closed circles represent data with values beyond the whiskers. Note, the y-axis for turbidity uses a base-10 log scale.

differences among sites and covariates. For example, the time smooth regression (i.e. the expected change in nitrate concentration) peaked around April-May and September-October at Arikaree, and in the cooler months at Lewis Run, whereas the pattern of peaks was less distinct at Caribou. The smooth regression of temperature (Fig 3) demonstrated a slight negative effect on the expected change in nitrate concentration at Arikaree. The reverse appeared to occur at Caribou and Lewis Run, with the expected change in nitrate concentration increasing with temperature.

For dissolved oxygen, the smooth regression indicated a minimal effect on expected change in nitrate concentration at Arikaree (Fig 3). This was also the case for Caribou up until a dissolved oxygen concentration of around 13.5 mg/L when the expected change in nitrate concentration increased sharply; however, this increase was due to a small number of high-concentration dissolved oxygen measurements only. The relationship between expected
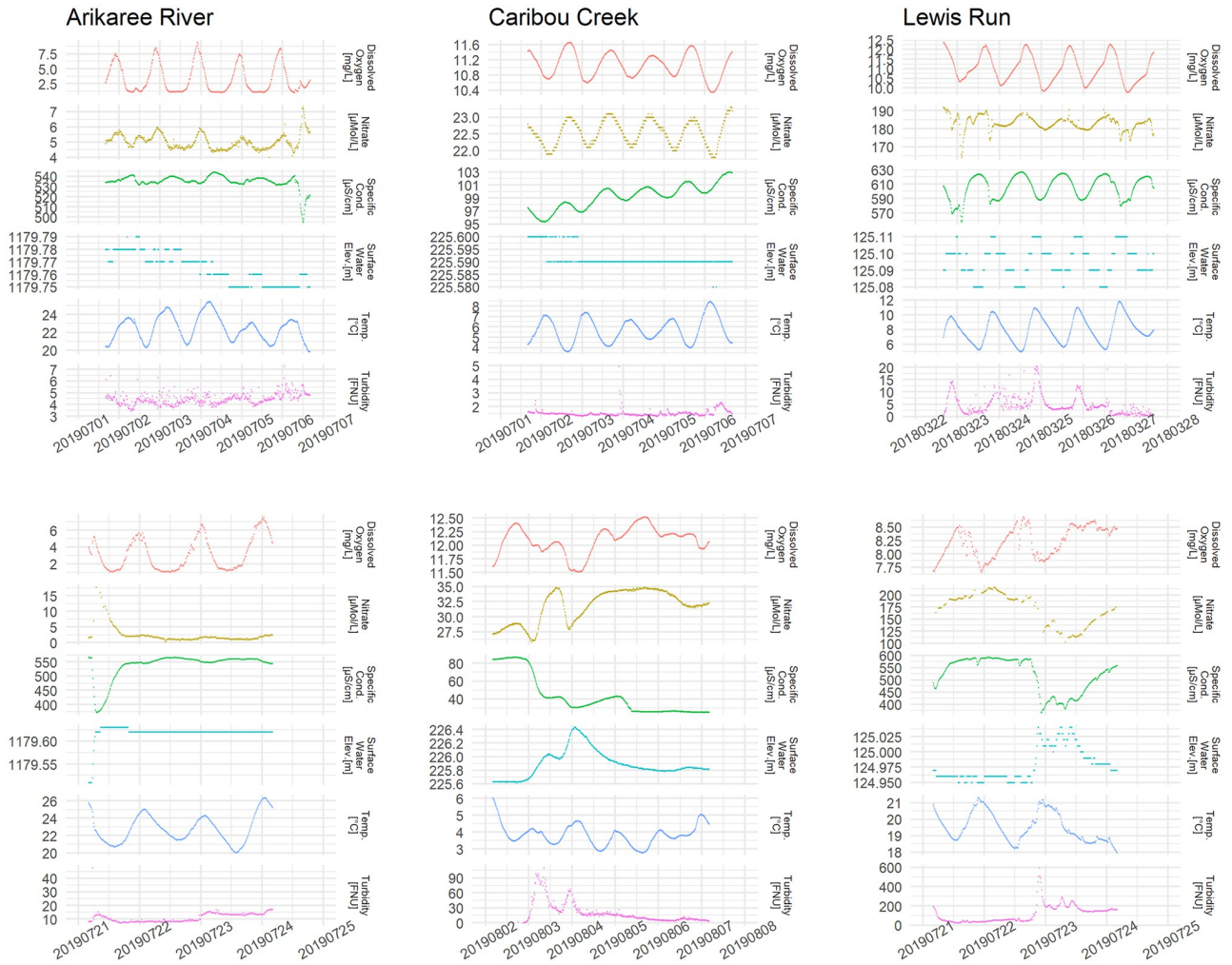
Fig 2. The top three plots are examples of diel fluctuations and other trends in water-quality at the study sides, visualised over 5-day windows of representative data, each starting and ending at 10 pm GMT. Local time in the three top plots starts at 3 pm at Arikaree, 2 pm at Caribou, and 5 pm at Lewis Run. The three bottom plots are examples of flow events, which are indicated by sudden rises in surface water elevation. Cond., conductance; Elev., elevation; Temp., temperature. Note that the scales for the y-axes differ among sites and variables.

https://doi.org/10.1371/journal.pone.0287640.g002

change in nitrate concentration and dissolved oxygen appeared to trough at Lewis Run around 9.5 mg/L.

The specific conductance smooth regression had a very different effect on the expected change in nitrate concentration at all three sites. At Caribou, there was a strong, negative effect on expected change in nitrate concentration up to around 45 μS/cm followed by a strong, positive effect to 65 μS/cm. Ranges of specific conductance at Arikaree and Lewis Run were similar, but smooth regressions were different for small values. At Arikaree, the expected change on nitrate concentration increased until 500μS/cm whereas it decreased at Lewis Run. In the two sites, when the specific conductance was higher than 550 μS/cm, nitrate concentration increased.

Confidence intervals for the smoothed regressions of log-transformed turbidity on the expected change in nitrate concentration tended to be wide in all three models compared with those of other covariates (Fig 3). The effect of turbidity, although weak, tended overall to be negative at Caribou and Arikaree, whereas at Lewis Run the effect tended to be more variable.
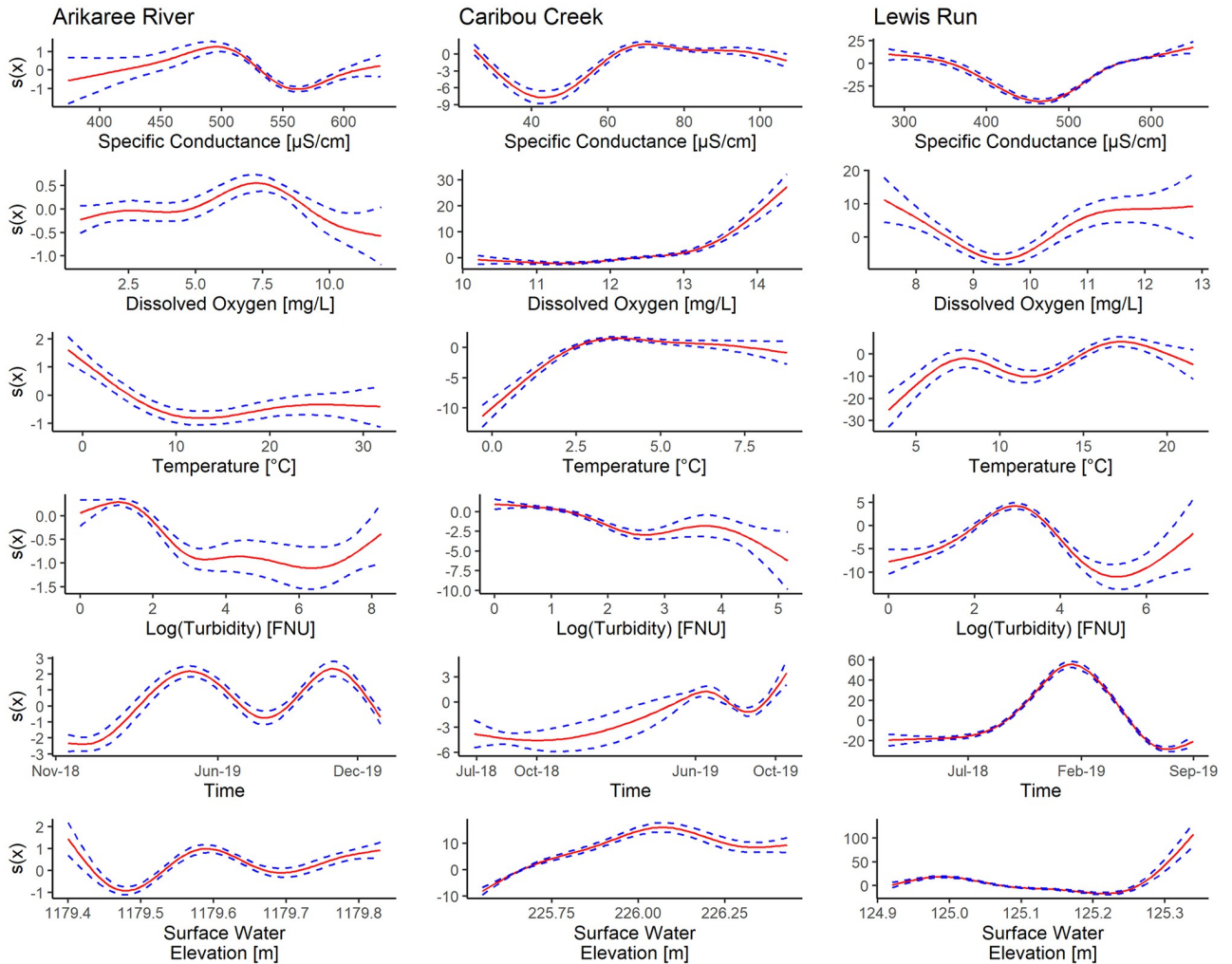
Fig 3. Smooth regressions (solid lines) of the expected change in nitrate concentration (y-axes) when values taken by covariates (x-axis) increased, by site. More specifically, each plot represents the regressive spline $s_k(z_{ki})$ (Eq (1)) of one covariate at one site, i.e. the expected change in nitrate concentration Y_i for a one unit increase in the covariate z_{ki}. Dashed lines show the standard error estimates. Missing values (gaps in the time series data) are not shown.

https://doi.org/10.1371/journal.pone.0287640.g003

The relationship between the smooth regression for surface water elevation and expected change in nitrate concentration was also different among sites. At Caribou, the expected change in nitrate concentration tended to increase with increasing surface water elevation, while Lewis Run exhibited an opposite relationship at least until near the end of the smooth regression. At Arikaree, confidence intervals were wide at the start of the smooth regression and cannot be interpreted, but at higher water stages, nitrate concentration tended to increase with surface water elevation, similar to Caribou.

GAMMs

The two-step modeling framework used in this study allowed us to differentiate the amount of deviance explained by covariates (GAM step) and the amount of deviance explained by the auto-regressive model (ARIMA step). In the first step, we included all variables as potential covariates in the GAMs because the VIF results indicated that multicollinearity among the

Table 3. Model performance for all three sites, as based on the approximated Akaike Information Criterion (aAIC).

Site	Model	aAIC
Arikaree	GAM	6044
	GAMM	-44631
Caribou	GAM	76871
	GAMM	40814
Lewis Run	GAM	10962
	GAMM	-13042

https://doi.org/10.1371/journal.pone.0287640.t003

covariates (the water-quality and time variables) was not a significant concern (VIFs < 6). The best GAMs, as selected by the stepwise variable-selection procedure based on the AIC, achieved a deviance explained of 75%, 83% and 85% respectively for Arikaree, Caribou, and Lewis Run The auto-regressive ARIMA functions then fit to the residuals of the best GAMs were ARIMA(4,0,0), ARIMA(5,0,3) and ARIMA(3,1,4), respectively, for Arikaree, Caribou and Lewis Run, which increased the total deviance explained by the models (see S2 File for fitted vs observed values). Hence, final GAMMs achieved a deviance explained of 99% for all three sites.

Model performance for each sites were evaluated using the aAIC (Eq 2). GAMMs performed far better than GAMs across all three sites (Table 3). However, the GAMs all explained a large proportion of variation in nitrate with the same combination of covariates, regardless of site. These included smooth terms for specific conductance, dissolved oxygen, temperature, turbidity (log-transformed), time and surface water elevation, which were all statistically significant (p < 0.001). However, the importance of covariates in the models fit to nitrate data differed among sites. Overall, the relatively low importance of all covariates at Arikaree compared to that at Caribou and Lewis Run (Fig 4) was congruent with the amount of deviance explained by the GAM portion of the Arikaree model (75%). The most important variable in the GAM for Arikaree was water temperature, but it explained less than 5% of the deviance in the final GAMM. However, the auto-regressive portion of the GAMM for Arikaree was particularly important, explaining 25% of the deviance. At Lewis Run, all the water-quality variables were important in explaining nitrate. The variable with the greatest importance was specific conductance (> 15%). The importance of all other water-quality covariates were between about 11% and 14% of the nitrate concentration deviance. These higher values of variable importance at Lewis Run corresponded with the larger proportion of deviance explained (85%) by the GAM part of the GAMM for this site. At Caribou, the auto-regressive part of the GAMM was less important in explaining nitrate concentration deviance (only 17%) than that explained by the GAM portion of the model (83%). Specific conductance was the most important variable (10%) and turbidity the least important in the Caribou model.

Discussion

Our study has demonstrated that GAMMs provide a suitable and useful method to model and understand the nonlinear relationships between nitrate and other water-quality variables, with an ability to explain 99% of the variation in nitrate concentration. Random Forests Regression (RFR) models, which can handle nonlinear interactions among variables and have an advantage over GAMM in their ability to provide information on variable importance, have been shown to explain 89% of nitrate concentration [21]. However, we note that [21] built their RFR models for the purpose of prediction and the resultant structure of their models was
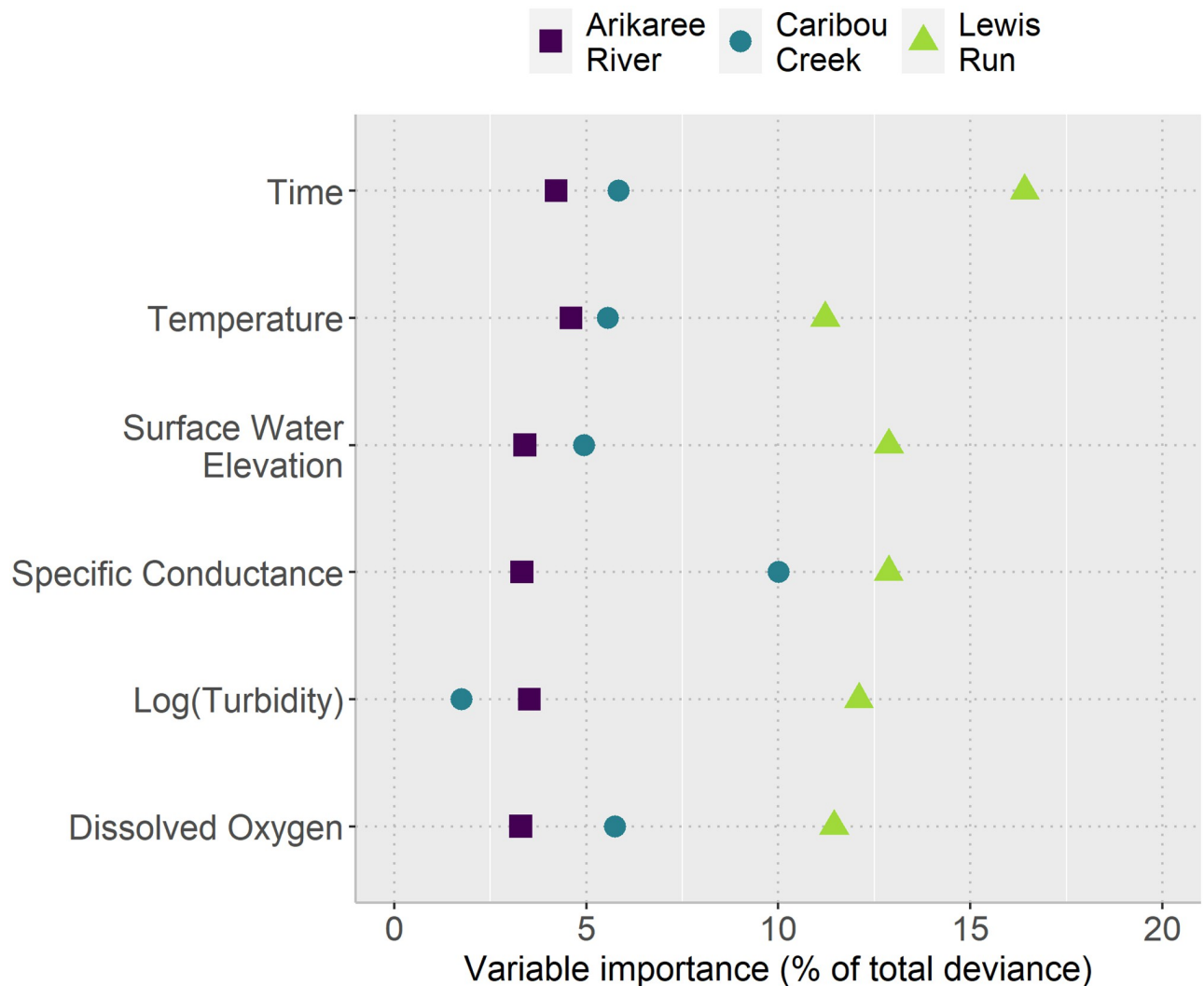
Fig 4. Variable importance, as the percentage of the total GAMM model deviance, for statistically significant covariates (p < 0.001), by study site. Total deviance explained by each GAMM was 99% for all three sites.

relatively opaque. Thus, we suggest GAMM regression in preference to other models like GLMMs and RFR when the primary aim is to explore the structure of the model and the shape of the non-linear relationship, which for the present study informs understanding of the links between nitrate and other water-quality variables. The two-step approach we used in this study is also highly relevant because, in addition to improving the understanding of the nitrate-water quality links, it enables the importance of temporal auto-correlation in nitrate measurements to be accounted for and assessed.

The use of GAMMs with high number of data has nevertheless raised computational challenges. The first challenge was the presence of missing data and technical anomalies in the time series. Within the two-year period of data analysed for this study, there were extended periods of such data that could not be incorporated into the models, which constrained the lengths of the time series modelled as a result (see S1 File). A second challenge was building the GAMMs themselves. The most commonly used function in the R mgcv package to fit GAMMs is 'gamm', which enables an ARIMA process to be fit to GAM residuals. However, as

explained in the gammmgcv vignette [43], 'gamm' is typically much slower to run than the 'gam' function, and the amount of memory required by R to run 'gamm' with large data sets may need to be increased substantially. Using the large data sets in this study often resulted in function failure, which required us to individually code the two-step GAMMs. This meant that temporal auto-correlation was not accounted for in the smoothing-parameter selection step, such that any auto-correlation could potentially be confused with trend resulting in under-estimation of the smoothing parameters and bias during inference. However, a two-step approach, such as that we used to secondarily account for auto-correlation, while typically not as efficient as estimating auto-correlation and smoothing parameters at the same time, is often more robust [44]. A final computational challenge in this study was to calculate variable importance for the GAMs. Several solutions are available to calculate variable importance for linear models (see, for example the `vimp` package [45]), but, to our knowledge, no solutions are available to easily calculate variable importance for non-linear models. As a result, we needed to determine variable importance using the relatively time-consuming method of iteratively comparing models with and without each covariate.

In terms of describing and better understanding nitrate dynamics and the relationship between nitrate concentration and other water-quality variables, the models we developed indicated similarities and differences among sites. Firstly, sites differed in their nitrate dynamics, likely relating to the distinct environmental conditions of the regions in which each sites was located. Rivers in regions with discontinuous permafrost, like Caribou, tend to export nitrogen (including in dissolved form, i.e. nitrate) rather than retain it, which is more typical of rivers in temperate regions like Arikaree and Lewis Run [46]. At Caribou, specific conductance was the most important variable explaining variation in nitrate concentration, followed by time to a lesser extent. Specific conductance was negatively correlated with nitrate and with surface water elevation (see S1 File). At Caribou, it is likely that rain storms are serving to increase stream level, decrease specific conductance and flush nitrate from shallow flow-paths through the watershed into the stream. Another hypothesis is that any link between specific conductance and nitrate concentration could be due to induced daily cycles of evapotranspiration [47], affecting water surface elevation, specific conductance, and nitrate dissolution in water. It is also possible that algae and other photosynthetic microorganisms active during the day were depleting nitrate. At Arikaree, no variable was especially important in explaining nitrate fluctuations. This suggests that daily fluctuations in nitrate at Arikaree were most likely due to the activity of algae and other photosynthetic microorganisms that use nitrate as an essential element for photosynthesis.

On the whole, water-quality variables at Caribou tended to reflect the relative stability of the subarctic Alaskan ecosystem. Caribou is located in a reserve upstream of urbanisation, with no known anthropogenic pollution present in chemical or physical form, including in the nitrate delivered into the system via atmospheric deposition (which has been monitored at the site since 1993 as part of the United States National Atmospheric Deposition Program). In contrast, multiple peaks and/or troughs in nitrate, turbidity and conductance occurred at Lewis Run, which may be associated with its proximity to an urban area, relative to other sites, and an upstream water treatment plant within a predominantly agricultural watershed. These factors would not only increase the overall nitrate concentration at Lewis Run, but could also cause temporal variability in water-quality [48]. In fact, the highest concentration of nitrate among the sites was observed at Lewis Run. The relatively high temporal variability in nitrate, turbidity and conductance at Lewis Run may also be affected by climate, being in a temperate zone with large intra-annual ranges in thermal and pluvial amplitudes. Water-quality at Arikaree also fluctuated substantially, and like Lewis Run, is located in an agricultural region, albeit within a semi-arid climate zone.

This work provides valuable insight on the links between nitrate and other water-quality variables in river systems, using approximately two years of data from each of three sites. While any analysis will only capture the range of variation contained within the input data itself, our choice of sites being distinct in their environmental and flow-regime characteristics, and the high-frequency nature of the sensor-based water-quality data that we analysed, means that inference can be drawn across a range of conditions. Continual analysis of data collected over future years, may, however, reveal new patterns and trends that allow for a more in-depth understanding of the relationships among variables. On the other hand the computational effort needed to create models from several years of data is large, and the availability of time series from high-frequency sensors that are absent from long or multiple sequences of data is limited. Nevertheless, testing the predictive ability of the models developed herein is likely to provide further insight on the links among water-quality variables while also enabling the prediction, for example, of missing or anomalous data in sensor-based time series.

Despite the among-site differences in nitrate concentration and other water-quality variables, the final GAMMs for each site included the same set of water-quality covariates. This demonstrates that these water-quality variables are consistently important for understanding variation in nitrate in rivers, even in watersheds with different types of land use and in different climate zones. The transferability of models, for example between different sites, remains a challenging obstacle in environmental and ecological modelling, as does the evaluation of their transferability [49]. However, our results suggested that a single model was not appropriate for the sites we examined, given the site-specific differences in relationship between nitrate and the other water-quality variables. Rather, we were able to identify a transferable modelling framework and a set of common of covariates that could together be used to explain nitrate concentration across disparate sites. Although GAMMs in such a framework must be tailored to data from individual sites, future research may reveal that models fit to data from sites with more similar land use, climate conditions and flow regimes are more transferable. With the implementation of automated sensing across several sites, watersheds and potentially regions, producing transferable models will become increasingly sought after in order to better understand water-quality relationships and dynamics, and to support water-resource management [13].

Conclusion

The findings of this study are highly relevant for scientists and managers responsible for *in-situ* monitoring in rivers. As mentioned above, gaps in *in-situ* sensor data are common and the methods demonstrated here could be applied to the problem of missing data imputation [50] and provide a more holistic description of nitrate dynamics. This is particularly important when financial resources are limited and decisions must be made about which sensors to buy and which water-quality variables to measure. In addition, this work provides a basis for future studies focused on the prediction of other critically important, water-quality variables that cannot be measured using *in-situ* sensors or when the sensors themselves are cost prohibitive.

Supporting information

S1 File. Original time-series. Time series of dissolved oxygen, nitrate concentration, specific conductance, surface water elevation, temperature, and turbidity in the three studied sites. (PDF)

S2 File. Diagnostic plots. Diagnostic plots of observed vs fitted values for the GAM and GAMM model constructed at each site. (PDF)

Acknowledgments

Author Contributions

Conceptualization: Benoit Liquet, Guy Litt, Kerrie Mengersen, Erin E. Peterson, Catherine Leigh.

Data curation: Rob J. Hyndman.

Formal analysis: Claire Kermorvant, Benoit Liquet, Guy Litt, Rob J. Hyndman.

Funding acquisition: Kerrie Mengersen.

Investigation: Benoit Liquet, Guy Litt, Catherine Leigh.

Methodology: Claire Kermorvant, Benoit Liquet, Erin E. Peterson, Rob J. Hyndman, Catherine Leigh.

Project administration: Erin E. Peterson, Catherine Leigh.

Resources: Guy Litt.

Supervision: Kerrie Mengersen, Erin E. Peterson, Rob J. Hyndman, Jeremy B. Jones, Jr., Catherine Leigh.

Validation: Kerrie Mengersen, Rob J. Hyndman, Jeremy B. Jones, Jr., Catherine Leigh.

Visualization: Claire Kermorvant.

Writing – original draft: Claire Kermorvant, Guy Litt, Catherine Leigh.

Writing – review & editing: Claire Kermorvant, Benoit Liquet, Guy Litt, Kerrie Mengersen, Erin E. Peterson, Rob J. Hyndman, Jeremy B. Jones, Jr., Catherine Leigh.

References

1. Camargo JA, Alonso A, Salamanca A. Nitrate toxicity to aquatic animals: A review with new data for freshwater invertebrates. Chemosphere. 2005; 58(9):1255–1267. https://doi.org/10.1016/j.chemosphere.2004.10.044 PMID: 15667845

2. Korostynska O, Mason A, Al-Shamma'a A. Monitoring of nitrates and phosphorus in wastewater: Current technologies and further challenges. International Journal on Smart Sensing and Intelligent Systems. 2012; 5(1). https://doi.org/10.21307/ijssis-2017-475

3. Boumans L, Fraters D, Van Drecht G. Nitrate leaching by atmospheric N deposition to upper groundwater in the sandy regions of the Netherlands in 1990. Environmental monitoring and assessment. 2004; 93(1-3):1–15. https://doi.org/10.1023/B:EMAS.0000016788.24386.be PMID: 15074606

4. Boesch DF, Boynton WR, Crowder LB, Diaz RJ, Howarth RW, Mee LD, et al. Nutrient enrichment drives Gulf of Mexico hypoxia. Eos, Transactions American Geophysical Union. 2009; 90(14):117–118. https://doi.org/10.1029/2009EO140001

5. Bricker SB, Longstaff B, Dennison W, Jones A, Boicourt K, Wicks C, et al. Effects of nutrient enrichment in the nation's estuaries: A decade of change. Harmful Algae. 2008; 8(1):21–32. https://doi.org/10.1016/j.hal.2008.08.028

6. Camargo JA, Ward J. Nitrate (NO3-N) toxicity to aquatic life: A proposal of safe concentrations for wo species of nearctic freshwater invertebrates. Chemosphere. 1995; 31(5):3211–3216. https://doi.org/10.1016/0045-6535(95)00182-8

7. Davidson J, Good C, Williams C, Summerfelt ST. Evaluating the chronic effects of nitrate on the health and performance of post-smolt Atlantic salmon *Salmo salar* in freshwater recirculation aquaculture systems. Aquacultural Engineering. 2017; 79:1–8. https://doi.org/10.1016/j.aquaeng.2017.08.003

8. Moore AP, Bringolf RB. Effects of nitrate on freshwater mussel glochidia attachment and metamorphosis success to the juvenile stage. Environmental pollution. 2018; 242:807–813. https://doi.org/10.1016/j.envpol.2018.07.047 PMID: 30032077

9. Lessels JS, Bishop TFA. Estimating water quality using linear mixed models with stream discharge and turbidity. Journal of Hydrology. 2013; 498:13–22. https://doi.org/10.1016/j.jhydrol.2013.06.006

10. Martinez K, Hart JK, Ong R. Environmental sensor networks. Computer. 2004; 37(8):50–56. https://doi.org/10.1109/MC.2004.91

11. Hart JK, Martinez K. Environmental Sensor Networks: A revolution in the earth system science? Earth-Science Reviews. 2006; 78(3):177–191. https://doi.org/10.1016/j.earscirev.2006.05.001

12. Jones AS, Horsburgh JS, Reeder SL, Ramirez M, Caraballo J. A data management and publication workflow for a large-scale, heterogeneous sensor network. Environmental monitoring and assessment. 2015; 187(6):348. https://doi.org/10.1007/s10661-015-4594-3 PMID: 25968554

13. Leigh C, Alsibai O, Hyndman RJ, Kandanaarachchi S, King OC, McGree JM, et al. A framework for automated anomaly detection in high frequency water-quality data from in situ sensors. Science of the Total Environment. 2019; 664:885–898. https://doi.org/10.1016/j.scitotenv.2019.02.085 PMID: 30769312

14. Park J, Kim KT, Lee WH. Recent advances in information and communications technology (ICT) and sensor technology for monitoring water quality. Water. 2020; 12(2):510. https://doi.org/10.3390/w12020510

15. Rodriguez-Perez J, Leigh C, Liquet B, Kermorvant C, Peterson E, Sous D, et al. Detecting technical anomalies in high-frequency water-quality data using Artificial Neural Networks. Environmental Science and Technology. 2020; 54(21):13719–13730. https://doi.org/10.1021/acs.est.0c04069 PMID: 32856893

16. Cawley KM. NEON Algorithm Theoretical Basis Document (ATBD): Water Quality NEON. DOC.004931vC; 2022. Available from: http://data.neonscience.org.

17. Ruddell BL, Zaslavsky I, Valentine D, Beran B, Piasecki M, Fu Q, et al. Sustainable long term scientific data publication: Lessons learned from a prototype Observatory Information System for the Illinois River Basin. Environmental Modelling and Software. 2014; 54:73–87. https://doi.org/10.1016/j.envsoft.2013.12.015

18. Leigh C, Kandanaarachchi S, McGree JM, Hyndman RJ, Alsibai O, Mengersen K, et al. Predicting sediment and nutrient concentrations from high-frequency water-quality data. PLOS ONE. 2019; 14(8):e0215503. https://doi.org/10.1371/journal.pone.0215503 PMID: 31469846

19. Chanat JG, Rice KC, Hornberger GM. Consistency of patterns in concentration-discharge plots. Water Resources Research. 2002; 38(8):22–1. https://doi.org/10.1029/2001WR000971

20. Diamantopoulou MJ, Papamichail DM, Antonopoulos VZ. The use of a neural network technique for the prediction of water quality parameters. Operational Research. 2005; 5(1):115–125. https://doi.org/10.1007/BF02944165

21. Harrison JW, Lucius MA, Farrell JL, Eichler LW, Relyea RA. Prediction of stream nitrogen and phosphorus concentrations from high-frequency sensors using Random Forests Regression. Science of the Total Environment. 2020; p. 143005. PMID: 33158521

22. Ensign SH, Doyle MW. Nutrient spiraling in streams and river networks. Journal of Geophysical Research: Biogeosciences. 2006; 111(G4):G04009. https://doi.org/10.1029/2005JG000114

23. Bowes MJ, Jarvie HP, Halliday SJ, Skeffington RA, Wade AJ, Loewenthal M, et al. Characterising phosphorus and nitrate inputs to a rural river using high-frequency concentration–flow relationships. Science of the Total Environment. 2015; 511:608–620. https://doi.org/10.1016/j.scitotenv.2014.12.086 PMID: 25596349

24. Wade AJ, Palmer-Felgate E, Halliday SJ, Skeffington RA, Loewenthal M, Jarvie H, et al. Hydrochemical processes in lowland rivers: Insights from in situ, high-resolution monitoring. Hydrology and Earth System Sciences. 2012; 16(11):4323–4342. https://doi.org/10.5194/hess-16-4323-2012

25. Dupas R, Jomaa S, Musolff A, Borchardt D, Rode M. Disentangling the influence of hydroclimatic patterns and agricultural management on river nitrate dynamics from sub-hourly to decadal time scales. Science of the Total Environment. 2016; 571:791–800. https://doi.org/10.1016/j.scitotenv.2016.07.053 PMID: 27422723

26. Vance J, Nance B, Monahan D, Mahal M, Cavileer M. NEON Preventive Maintenance Procedure: AIS Surface Water Quality Multisonde. NEON.DOC.001569vB; 2019. Available from: https://data.neonscience.org/data-products/DP1.20288.001/RELEASE-2021.

27. Willingham R, Cavileer M, Csavina J, Monahan D. NEON Preventive Maintenance Procedure: Submersible Ultraviolet Nitrate Analyzer. NEON.DOC.002716vB; 2019. Available from: http://data.neonscience.org.

28. Cawley KM. NEON Aquatic Sampling Strategy NEON.DOC001152vA; 2016. Available from: http://data.neonscience.org.

29. NEON (National Ecological Observatory Network). Nitrate in surface water (DP1.20033.001), RELEASE-2021. https://doi.org/10.48443/924T-1K41. Available from: https://data.neonscience.org/data-products/DP1.20033.001/RELEASE-2021.

30. NEON (National Ecological Observatory Network). Water quality (DP1.20288.001, RELEASE-2021. https://doi.org/10.48443/D8KW-5J62). Available from: https://data.neonscience.org/data-products/DP1.20288.001/RELEASE-2021.

31. NEON (National Ecological Observatory Network). Temperature (PRT) in surface water (DP1.20053.001), RELEASE-2021. https://doi.org/10.48443/NY19-PJ91. Available from: https://data.neonscience.org/data-products/DP1.20053.001/RELEASE-2021.

32. NEON (National Ecological Observatory Network). Elevation of surface water (DP1.20016.001), RELEASE-2021. https://doi.org/10.48443/QSER-8M94. Available from: https://data.neonscience.org/data-products/DP1.20016.001/RELEASE-2021.

33. Kroll CN, Song P. Impact of multicollinearity on small sample hydrologic regression models. Water Resources Research. 2013; 49(6):3756–3769. https://doi.org/10.1002/wrcr.20315

34. Helsel DR, Hirsch RM. Statistical Methods in Water Resources. Elsevier; 1992.

35. Hastie T, Tibshirani R. Generalized Additive Models. CRC press; 1990.

36. Wood SN. Generalized Additive Models: An Introduction with R. CRC press; 2017.

37. Akaike H. A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. Biometrika. 1979; 66(2):237–242. https://doi.org/10.1093/biomet/66.2.237

38. Faraway JJ. Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models. 2nd ed. Chapman and Hall/CRC; 2016.

39. Burnham KP, Anderson DR. Model Selection and Multimodel Inference: A Practical Information-theoretic Approach. Springer-Verlag; 2002.

40. Fox J, Weisberg S, Adler D, Bates D, Baud-Bovy G, Bolker B, et al. car: Companion to Applied Regression; 2020. Available from: https://CRAN.R-project.org/package=car.

41. Hastie T. gam: Generalized Additive Models; 2020. Available from: https://CRAN.R-project.org/package=gam.

42. Hyndman R, Athanasopoulos G, Bergmeir C, Caceres G, Chhay L, O'Hara-Wild M, et al. forecast: Forecasting functions for time series and linear models; 2021. Available from: https://CRAN.R-project.org/package=forecast.

43. Wood SN. gammmgcv: Generalized Additive Mixed Models; 2021. https://rdrr.io/cran/mgcv/man/gamm.html.

44. Casella G, Berger RL. Statistical Inference. Pacific Grove, CA: Duxbury; 2002.

45. Williamson B, Gilbert P, Carone M, Simon R. Nonparametric variable importance assessment using machine learning techniques. Biometrics. 2020.

46. Jones JB Jr, Petrone KC, Finlay JC, Hinzman LD, Bolton WR. Nitrogen loss from watersheds of interior Alaska underlain with discontinuous permafrost. Geophysical Research Letters. 2005; 32(2).

47. Yashi DK, Suzuki K, Nomura M. Diurnal fluctuation in stream flow and in specific electric conductance during drought periods. Journal of Hydrology. 1990; 115(1):105–114. https://doi.org/10.1016/0022-1694(90)90200-H

48. Blann KL, Anderson JL, Sands GR, Vondracek B. Effects of agricultural drainage on aquatic ecosystems: A review. Critical reviews in environmental science and technology. 2009; 39(11):909–1001. https://doi.org/10.1080/10643380801977966

49. Yates KL, Bouchet PJ, Caley MJ, Mengersen K, Randin CF, Parnell S, et al. Outstanding challenges in the transferability of ecological models. Trends in Ecology and Evolution. 2018; 33(10):790–802. https://doi.org/10.1016/j.tree.2018.08.001 PMID: 30166069

50. Kermorvant C, Liquet B, Litt G, Jones JB, Mengersen K, Peterson EE, et al. Reconstructing missing and anomalous data collected from high-frequency in-situ sensors in fresh waters. International Journal of Environmental Research and Public Health. 2021; 18(23). https://doi.org/10.3390/ijerph182312803 PMID: 34886529