Supplementary materials for this article are available at https://doi.org/10.1007/s13253-022-00513-2.

Model-Based Clustering of Trends and Cycles of Nitrate Concentrations in Rivers Across France

Matthew HEINER, Matthew J. HEATON, Benjamin ABBOTT, Philip WHITE, Camille MINAUDO, and Rémi DUPAS

Elevated nitrate from human activity causes ecosystem and economic harm globally. The factors that control the spatiotemporal dynamics of riverine nitrate concentration remain difficult to describe and predict. We analyzed nitrate concentration from 4450 sites throughout France to group sites that exhibit similar trend and seasonal behaviors during 2010–2017 and relate these dynamics to catchment characteristics. We employed a latent-variable, Bayesian mixture of harmonic regressions model to infer site clustering based on multi-year trend and annual cycle amplitude and phase. We examined clustering patterns and relationships among nitrate level, trend, and seasonality parameters. Cluster membership probabilities were governed by continuous, latent variables that were informed with seven classes of covariates encompassing geology, hydrology, and land use. To relate interpretable parameters to the covariates, we modeled amplitude and phase separately in a novel framework employing a bivariate phase regression with the projected normal distribution. The analysis identified regional regimes of nitrate dynamics, including trend classifications. This approach can reveal general patterns that transcend small-scale heterogeneity, complementing site-level assessments to inform regional- to national-level progress in water quality.

Supplementary materials accompanying this paper appear on-line.

Key Words: Directional data; Harmonic regression; Hierarchical models; Hydrology; Mixture modeling.

M. Heiner (✉) · M. J. Heaton · P. White, Department of Statistics, Brigham Young University, Provo, USA(E-mail: *heiner@stat.byu.edu*).

B. Abbott, Department of Plant and Wildlife Sciences, Brigham Young University, Provo, USA.

C. Minaudo, Physics of Aquatic Systems Laboratory, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.

R. Dupas, INRAe, L'institut Agro, UMR 1069 SAS Rennes, France.

Journal of Agricultural, Biological, and Environmental Statistics, Volume 28, Number 1, Pages 74–98
https://doi.org/10.1007/s13253-022-00513-2

1. INTRODUCTION

1.1. PROBLEM BACKGROUND

Excess nutrients from agriculture, wastewater, and fossil fuel use have caused harmful algal blooms and hypoxic dead zones in two-thirds of aquatic ecosystems worldwide (Diaz and Rosenberg 2008; Frei et al. 2020). This process of nutrient enrichment, referred to in the scientific literature as eutrophication (Le Moal et al. 2019), degrades health and water security globally, causing trillions in USD of economic damage annually (Dupas et al. 2019; Cheng et al. 2020). Despite substantial investments in nutrient monitoring and mitigation by governments since the mid-20th century, eutrophication has proven extremely persistent in developed and developing countries (Conley et al. 2009; Stoddard et al. 2016; Osgood 2017; Le Moal et al. 2019; Hannah et al. 2022).

One of the challenges to solving eutrophication is that nutrient concentrations show high levels of spatiotemporal variation in both surface and groundwater environments (Bochet et al. 2020; Kolbe et al. 2019; Abbott et al. 2018b). Changes in water flow, nutrient source, and biological activity can trigger large shifts in nutrient concentration on hourly to centennial timescales (Moatar et al. 2017; Messer et al. 2019; Ascott et al. 2021). This complicates finding nutrient sources and quantifying long-term trends such as improvement or degradation following changes in management or environmental conditions (Minaudo et al. 2019; Smits et al. 2019; Ehrhardt et al. 2019). To characterize this variability, researchers and regulatory agencies measure nutrient concentrations throughout the year at multiple locations (Abbott et al. 2018a; Moatar et al. 2020). This time- and resource-intensive monitoring is necessary to assess overall ecological status and demonstrate compliance with environmental legislation (Zhang et al. 2021).

The growing global database of spatially extensive and long-term water quality monitoring presents a challenge and opportunity for the mathematical and statistical sciences. New approaches are needed to generate understanding of ecological processes and to develop ideal monitoring frameworks from large, multi-dimensional water databases (Hartmann et al. 2014; Zarnetske et al. 2018). For example, consider Fig. 1, which displays nitrate (NO_3) concentration at five river monitoring sites in France from January 1, 2010, to December 31, 2016. Just among these several rivers, there are large differences in sampling frequency, seasonal variability, and overall NO_3 concentration. These irregularities are created by ecological diversity of the systems as well as inconsistency in monitoring approaches, where field access and funding may themselves vary on seasonal to decadal timescales (Burt and McDonnell 2015; Jiang et al. 2020). Considering that similar data are collected at more than 5000 stations in France alone (Dupas et al. 2019), developing tools to characterize and interpret nutrient variability and central tendency are greatly needed.

Various statistical tools have been applied to characterize complex hydrochemical time series, including self-organizing maps, fractal scaling methods, and frequency decomposition (Chiverton et al. 2015; Kirchner and Neal 2013; Lloyd et al. 2014; Underwood et al. 2017). While each of these families of methods has pros and cons, irregularities in data sources and difficulty interpreting model outputs have been perennial problems. Without accounting for interrelated variation on seasonal to inter-annual timescales, short- and long-

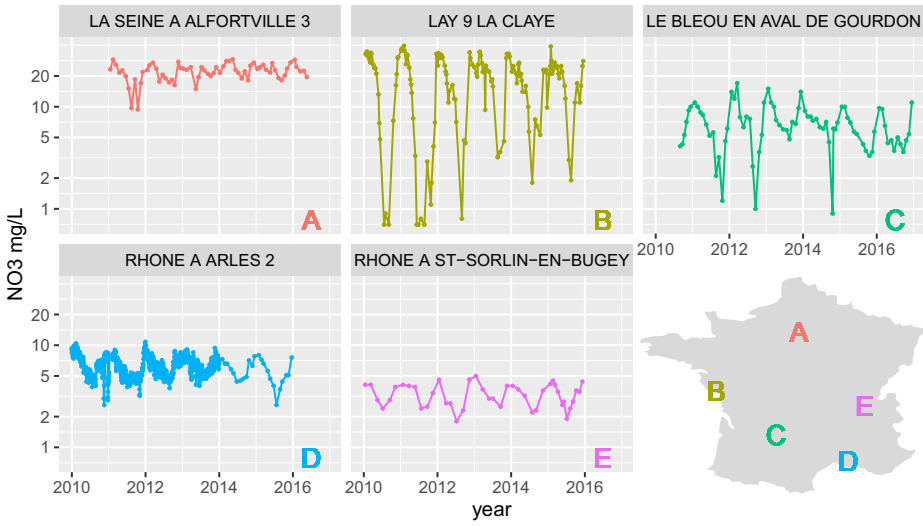

Figure 1. Time series of nitrate (NO_3) concentrations (in mg/L, reported on a logarithmic scale) for five stations in France.

term metrics of nutrient status could be obscured. Even simple parameters such as the linear trend, amplitude, phase, and (residual) standard deviation of time series can vary widely at regional to national scales in non-independent ways (Fig. 2). Adequately describing temporal variability of diverse nutrient time series is necessary to identify what spatial differences (e.g., land management, climate, topography, or vegetation) account for nutrient release or retention (Frei et al. 2020; Minaudo et al. 2019).

In this article, we focus on two of the most policy-relevant parameters in NO_3 time series: seasonality and long-term trend. The amplitude of seasonal change in NO_3 can influence whether a site surpasses regulatory limits and indicate whether a watershed is still being overloaded with nutrients (Abbott et al. 2018b; Newcomer et al. 2021; Ebeling et al. 2021). Seasonal phase is indicative of possible coupling with hydrological variations and timing of activation of sources and retention processes (Guillemot et al. 2020; Vaughan et al. 2017). Accurate descriptions of multi-annual trend is particularly important to assess effectiveness of nutrient mitigation efforts (Dupas et al. 2018; Frei et al. 2020).

1.2. Research Goals and Approach

The goals of this work are to (i) estimate the seasonal to inter-annual patterns in log-NO_3 concentrations for all sites in the French river monitoring network and (ii) investigate how environmental conditions influence these patterns to understand what factors may drive nutrient release and associated eutrophication. Regarding the first goal, from a statistical modeling perspective, we face a few issues. First, the infrequent sampling of concentration at some sites results in few data points available for statistical inference of temporal characteristics. To accomplish our goal in spite of this challenge, we borrow information across similar rivers to estimate parameters. This can be done via traditional geostatistical meth-

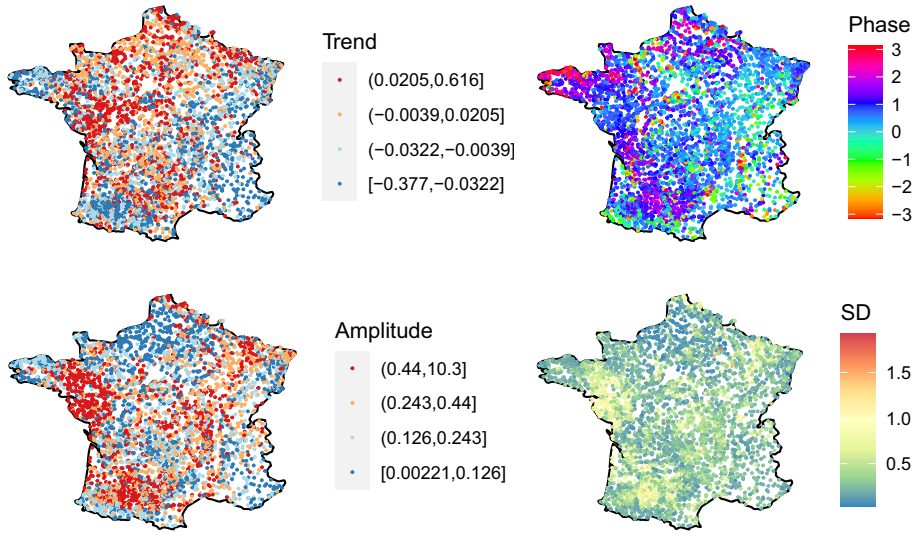

Figure 2. Maps of temporal trend (top left), phase (top right) and amplitude (bottom left) of the annual cycle, and residual standard deviation (bottom right) for independently fitted time series of log-NO_3 concentration, reported by station (Color figure online).

ods (Cressie 2015; Banerjee et al. 2014). However, as discussed by Garreta et al. (2010), Ver Hoef and Peterson (2010), and Isaak et al. (2014), such methods are not generally suitable for modeling river networks due to spatial dependence that is not a function of Euclidean distance. Alternatively, we could adopt an approach similar to Álvarez-Cabria et al. (2016), de Lavenne et al. (2016), Zimmerman and Ver Hoef (2017), or O'Donnell et al. (2014) and build directional spatial dependence based on routing within river networks (Pearse et al. 2020). However, this information is not always available or reliable at large scales due to inconsistencies in spatial data formats and subsurface hydrological complexity not associated with surface topography (Schaller and Fan 2009; Yan et al. 2019).

In this research, we employ clustering to borrow information across rivers in estimating seasonality and trend. Clustering approaches have been used successfully in the analysis of river networks, by, among others, Kim and Seo (2015), de Almeida et al. (2019), and Zubaidah et al. (2018). The advantage of the clustering approach is that clusters need not be defined spatially and can capture the heterogeneous spatial relationships in temporal characteristics we see in Fig. 2. By inferring latent clusters for the river network, we circumvent errors in river network delineation to identify hydrochemical differences that create distinct temporal patterns.

Our second goal is to characterize relationships between the temporal characteristics of the rivers with information about the contributing watersheds. Specifically, in addition to the NO_3 time series, we have external covariates for each river monitoring site that describe local geography, hydrology, and land use. The challenge is that these covariates are temporally static, as is typical for large-scale spatial data. Further, because cluster-specific parameters are not identifiable due to label switching (Stephens 2000; Jasra et al. 2005; Sperrin et al.

2010; Rodríguez and Walker 2014), we cannot directly relate the available covariates to such parameters. As a solution to these issues, we devise a modeling strategy by enforcing ordering constraints on the cluster labels. Under such ordering constraints, we show that relating the covariates to the parameters is equivalent to a linear regression with an ordinal categorical response.

An additional challenge associated with our second goal is the fact that one of the temporal parameters (phase) is a circular parameter. Strategies for doing regression with a circular response typically involve wrapped distributions (Ravindran and Ghosh 2011), the projected normal distribution (Nuñez-Antonio and Gutiérrez-Peña 2005; Nuñez-Antonio et al. 2011), or the generalized projected normal distribution (Wang and Gelfand 2013). However, again due to the clustering approach taken above, our situation is slightly different than these approaches in that we have a discrete set of circular parameters (one for each cluster). In a novel piece of modeling, we develop a latent variable approach adapted from Albert and Chib (1993), Higgs and Hoeting (2010), Berrett and Calder (2012), Schliep and Hoeting (2013), and Berrett and Calder (2016), that uses the projected normal distribution to model a discrete, ordinal, circular parameter as a function of covariates.

The French monitoring network data used in this research are provided by water authorities and are publicly available (Naïades; http://www.naiades.eaufrance.fr/france-entiere#/). These data are used for regulatory purposes (per the European Union Water Framework Directive), with common laboratory protocols and standards throughout Europe, and are therefore meant to be comparable in space and time.

Our analysis integrates 246,189 observations from 4450 monitoring sites, referred to as stations, throughout 2010–2016. Station-specific time series generally contain between 30 and 80 (median 54) observations, with only 79 of 4450 stations exceeding 80. This amount of data, combined with the fact that observations within a station are temporally correlated, presents a computational challenge. In performing estimation of model parameters, we need to ensure that such estimation is done efficiently. For our modeling framework, we augment the parameter space with latent variables and show that the subsequent complete conditional distributions of most parameters are conjugate with respect to their prior distribution, thereby facilitating estimation via Gibbs sampling. For those parameters that are not conjugate, we employ other efficient techniques described in Sect. 2 and the Supplementary Materials.

The remainder of this paper is outlined as follows: In Sect. 2, we describe our modeling approach along with the computational strategy for parameter estimation. Section 3 fits the proposed model to the France river data, and Sect. 4 draws conclusions and highlights areas for future statistical and ecological research.

2. A CLUSTERED RIVER MODEL

Let $y_r(t)$ denote the NO_3 concentration at time $t \in [0, T]$ in river station $r = 1, \ldots, R$, where time t has been scaled to units of years, with $t = 0$ corresponding to January 1, 2010, and $t = T$ corresponding to December 31, 2016. The natural logarithm of NO_3 concentration is amenable to a simpler representation of the annual cycle and admits far more convenient modeling and computation than the nonnegative, positive-skewed raw concentration. The

basic model for each time series is

$$\log(y_r(t)) = \delta_r + \beta_{Z_\beta(r)} t + \alpha_{Z_\alpha(r)} \cos\left(2\pi t - \phi_{Z_\phi(r)}\right) + \epsilon_r(t) \tag{1}$$

where δ_r is a station-specific intercept, $\beta \in \mathbb{R}^1$ corresponds to a linear temporal trend, $\alpha > 0$ corresponds to a temporal amplitude, $\phi \in (-\pi, \pi)$ corresponds to a temporal phase shift and $(\epsilon_r(t))$ is model error. The $Z_\beta(r)$, $Z_\alpha(r)$ and $Z_\phi(r)$ are latent indicator mappings where, for example, $Z_\beta(r) : \{1, \ldots, R\} \to \{1, \ldots, K_\beta\}$ maps a station r onto a latent cluster membership $\{1, \ldots, K_\beta\}$ for annual temporal trend and K_β corresponds to the number of latent clusters. We define $Z_\alpha(r)$ and $Z_\phi(r)$ similarly for the amplitude and phase clusters with K_α and K_ϕ corresponding to the number of latent amplitude and phase clusters, respectively.

Intuitively, (1) defines three sets of station clusters: one based on each of temporal trend, amplitude, and phase. That is, two stations can be clustered together based on one characteristic of the NO_3 time series while not clustered based on other characteristics. This three-set approach to clustering is important because such clustering allows for strong borrowing of information to estimate the temporal trends, amplitudes, and phases while adding flexibility for one station to borrow information from different sets of stations that most closely resemble its own temporal characteristics. Note also that (1) employs a less convenient (nonlinear) parameterization of the annual cycle in order to preserve both separate and interpretable clustering of the amplitude and phase.

Although (1) is unrealistically simple as a model for complex dynamics of NO_3 concentration, it successfully discriminates along four primary modes of variation in the time series. With data from 4450 sites across France, we prioritize both interpretability and computational scalability. Together with three-way clustering, the model in (1) balances our inferential goals and these criteria.

As discussed in the introduction, spatial correlation in river systems is not driven by spatial proximity (Euclidean distance) but rather by the network topology (Pearse et al. 2020). Additionally, explicit use of geostatistical models can greatly add computation time and cost. Because network topology information is not available for this research and the additional model details below are already computationally demanding, we assume spatial independence among stations conditional on cluster parameters. While we make the simplifying assumption of spatial independence, we do assume temporal correlations within a station. Specifically, we assume $\epsilon_r(t)$ have temporal covariance $\text{Cov}(\epsilon_r(t), \epsilon_r(t')) = \sigma_r^2 \exp\{-d_r|t - t'|\}$ where d_r is a decay parameter for station r and σ_r^2 is the marginal variance.

To prevent arbitrary label switching, we require ordering within the trend and amplitude parameter vectors, i.e., $\beta_1 < \beta_2 < \cdots < \beta_{K_\beta}$ and $\alpha_1 < \alpha_2 < \cdots < \alpha_{K_\alpha}$. As a consequence, the latent indicators $Z_\beta(r)$ and $Z_\alpha(r)$ are *ordered* multinomial random variables. Under this ordering, we can further employ latent Gaussian models of Albert and Chib (1993) for Z_β and Z_α. That is, for $j \in \{\beta, \alpha\}$, we can set

$$Z_j(r) = \sum_{k=1}^{K_j} k \, \mathbb{1}\{c_{j(k-1)} < U_{jr} < c_{jk}\} \tag{2}$$

where $\mathbb{1}\{\cdot\}$ is an indicator variable, $c_{j0} = -\infty < c_{j1} < \cdots < c_{jK_j} = \infty$ are cut points and $U_{jr} \sim \mathcal{N}(\boldsymbol{x}_r'\boldsymbol{\theta}_j, 1)$ are latent Gaussian random variables, $\boldsymbol{x}_r = (1, x_{r1}, \ldots, x_{rP})'$ is a vector of covariates specific to station r with corresponding coefficients $\boldsymbol{\theta}_j = (\theta_{j0}, \ldots, \theta_{jP})'$. Under this model specification, the variables in $\boldsymbol{x}_r$ explain the probability that each station belongs to a certain cluster. For example, if $\theta_{jp} > 0$, then as x_{rp} increases, station r is more likely to belong to a cluster with a larger linear time trend (if $j = \beta$) or amplitude (if $j = \alpha$). Importantly, this allows us to infer how characteristics of the river (such as location, elevation or other explanatory variables) influence the temporal traits of NO_3 at the station. The covariates in $\boldsymbol{x}_r$ are discussed in Sect. 3.3.

While the above latent, Gaussian framework is effective for Z_β and Z_α, the phase variables $\phi_1 < \cdots < \phi_{K_\phi}$ are ordered circular random variables. That is, a phase shift of $\phi \approx \pi$ is effectively the same as a phase shift of $\phi \approx -\pi$. To account for the circular nature of phase, we model

$$Z_\phi(r) = \sum_{k=1}^{K_\phi} k\, \mathbb{1}\{c_{\phi(k-1)} < U_{\phi r} < c_{\phi k}\} \tag{3}$$

where $c_{\phi 0} = -\pi < c_{\phi 1} < \cdots < c_{\phi K_\phi} = \pi$ are cut points that partition the interval $(-\pi, \pi)$ and $U_{\phi r} \sim \mathcal{PN}(\boldsymbol{\mu}_{\phi r}, \boldsymbol{I})$ where $\mathcal{PN}(\boldsymbol{\mu}, \boldsymbol{I})$ is the projected normal distribution with mean direction $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{I}$. That is, if we define a bivariate latent vector $\boldsymbol{D}_{\phi r} = (D_{\phi r1}, D_{\phi r2})' \sim \mathcal{N}(\boldsymbol{\mu}_\phi(r), \boldsymbol{I})$, then $U_{\phi r} = \text{atan2}(D_{\phi r2}, D_{\phi r1}) \sim \mathcal{PN}(\boldsymbol{\mu}_\phi(r), \boldsymbol{I})$ so that $U_{\phi r}$ is the angle between the origin and the vector $\boldsymbol{D}_{\phi r}$. Importantly, the projected normal distribution with identity covariance is rotationally symmetric about the mean direction $\boldsymbol{\mu}$. Practically, this allows circular probabilities for $Z_\phi(r)$ in that both $\text{Prob}(Z_\phi(r) = 1)$ and $\text{Prob}(Z_\phi(r) = K_\phi)$ can be large simultaneously, which would not be possible under a standard Gaussian distribution. The concept of wrapped probabilities is illustrated in Fig. 3, which shows the $\mathcal{PN}(\boldsymbol{\mu}, \boldsymbol{I})$ distribution with equally spaced cut points on $[-\pi, \pi]$ with two different choices for $\boldsymbol{\mu}$.

In a similar model setup as above, we wish to explain phase cluster membership with station-specific covariates. Following Nuñez-Antonio and Gutiérrez-Peña (2005) and Nuñez-Antonio et al. (2011), to explain phase cluster membership we can set $\boldsymbol{\mu}_\phi(r) = \boldsymbol{\Theta}_\phi \boldsymbol{x}_r$ where $\boldsymbol{\Theta}_\phi = \{\theta_{\phi ip}\}_{ip}$ is a $2 \times (P+1)$ matrix of coefficients associated with the station-specific covariate vector $\boldsymbol{x}_r$. Due to the circular nature of phase, the interpretation of each $\theta_{\phi ip}$ is not as straightforward as $\theta_{\beta p}$ and $\theta_{\alpha p}$, so we impose additional constraints. Specifically, we set $K_\phi = 12$ and *a priori* constrain $c_{\phi(k-1)} < \phi_k < c_{\phi k}$ where the 12 phase cut points are fixed at equally spaced intervals that align with the (D_1, D_2) axes. This endows the four quadrants of the unconstrained $\mathbb{R}^2$ space for $\boldsymbol{D}_\phi$ to correspond with (approximate) three-month seasons and each ϕ_k with an associated month. The mapping from $\mathbb{R}^2$ to the interval $(-\pi, \pi)$ is depicted in the left panel of Fig. 3, with two bivariate normal density contours for illustration. The resulting projected normal densities for U_ϕ determining cluster membership probabilities are shown on the right. Although cluster representative ϕ_k parameters may vary within their respective intervals, the cardinal directions maintain a general interpretation. That is, positive values of $\theta_{\phi 1 p}$ indicate that a positive covariate value pushes $\boldsymbol{\mu}_\phi$ to the right (positive D_1 direction) and approximately coincides with increased

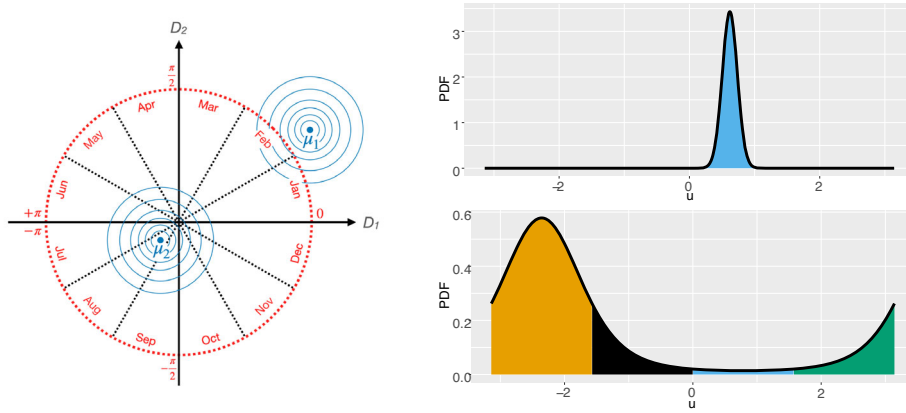

Figure 3. Illustration of mapping probability density in the unconstrained $\boldsymbol{D}_\phi$ space (left) to wrapped probabilities for phase cluster membership (right). Two example values of $\boldsymbol{\mu}_\phi = \boldsymbol{\Theta}_\phi \boldsymbol{x}$ are shown: $\boldsymbol{\mu}_1 = (7, 5)$ and $\boldsymbol{\mu}_2 = (-1, -1)$, together with contours tracing the corresponding bivariate densities. The angle from the positive D_1 axis is superimposed, indicating angular cut points and approximate month labels. The probability densities for $U_{\phi 1}$ and $U_{\phi 2}$, corresponding to $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, are on top and bottom, respectively. Shading indicates the probabilities of falling within the four primary seasonal divisions (Color figure online).

probability of falling into a cluster represented after the autumnal equinox and before the vernal equinox (in France). Positive values of $\theta_{\phi 2p}$ with positive covariate values increase the probability of falling into clusters represented approximately after the winter solstice and before the summer solstice. Thus, if both $\theta_{\phi 1p}$ and $\theta_{\phi 2p}$ are positive, a positive covariate value would push the corresponding $\boldsymbol{\mu}_\phi$ toward the first quadrant, as with $\boldsymbol{\mu}_1$ in Fig. 3.

Inference for the above model was carried out via Bayesian methodology (Reich and Ghosh 2019). Details of the posterior sampling algorithm are found in Supplementary Materials, but we highlight a few pieces of the computation here. First, using principles of parameter expansion, we can sample the θ parameters directly from their complete conditional distribution. Second, when considering a sampling approach for the station-specific cluster indicators, note that (2) and (3) provide a non-stochastic relationship between U and Z. We sample both U and Z via composition, as detailed in Cowles (1996). Third, we employ slice sampling on phase parameters (ϕ_k) and cut points ($\{c_{\beta k}\}_{k=2}^{K_\beta - 1}$ and $\{c_{\alpha k}\}_{k=2}^{K_\alpha - 1}$; Neal, 2003). Finally, to help maintain cluster identity and ensure all clusters are populated, two stations are assigned fixed membership in each trend and amplitude cluster, and are not updated (Kunkel and Peruggia 2020). These sites are selected during model initialization as those whose preliminary estimates are closest to initial cluster-center values.

3. APPLICATION TO RIVERS IN FRANCE

We fit the proposed model at two resolutions by varying the number of trend and amplitude clusters while holding $K_\phi = 12$ fixed. The (K_β, K_α) pairs used were (5,7) and (7,11). In both models, we fixed the center trend parameter (the third in the $K_\beta = 5$ model and the fourth in the $K_\beta = 7$ model) at zero to furnish a testable hypothesis of no trend. Although

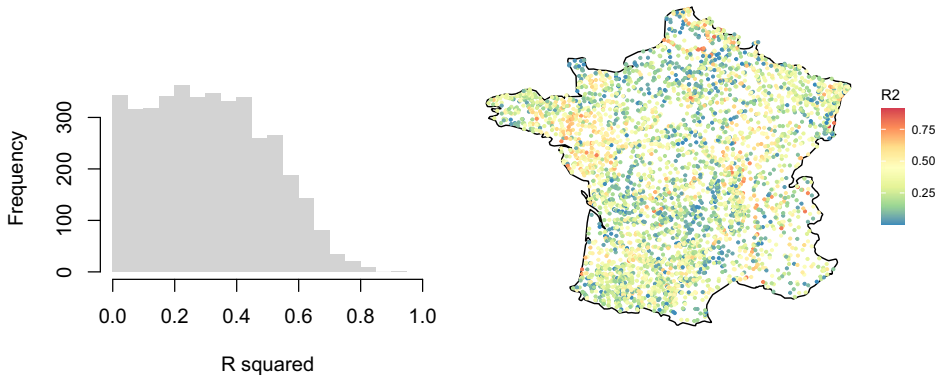

Figure 4. Histogram (left) and spatial distribution (right) of posterior mean R-squared values from the (5,7) model fit. Negative values of R-squared are excluded (Color figure online).

we inspect goodness-of-fit metrics to aid in selecting among the three models, our primary objective is to achieve balance between flexibility and interpretability.

The MCMC sampler for the (5,7) model was run for 66,000 iterations, with the first 42,000 discarded as burn-in, and the remaining iterations thinned to 2400 inference samples (88,000 total iterations were used for the (7,11) model). Four such chains were initialized with parameter values near those found by warm-up runs initialized with a K-means algorithm. Parameters of inferential interest indicate adequate mixing and stable results in the (5,7) model, with exception of a few trend coefficients (and associated cluster memberships), which exhibit drift and vary on long timescales. We believe this is due to some estimated trends being small and statistically indistinguishable. While there is evidence for fewer trend clusters, we elected to leave $K_\beta = 5$ to allow for a cluster with $\beta = 0$ fixed and two levels each of positive and negative departures. We explore MCMC diagnostics in the Supplementary Materials.

3.1. Model Assessment and Selection

The mixture modeling literature is rich with methods for selecting the number of components/clusters; perhaps most notable are predictive criteria , in or out of sample. We approach the problem from multiple angles, attempting to balance model goodness of fit, interpretability, and cluster similarity. Additional details for each of the approaches are given in the Supplementary Materials.

Pareto-smoothed leave-one-out cross-validation and WAIC (Vehtari et al. 2017) tentatively favor the (5,7) model over the (7,11) alternative, although the information criterion values are well within one standard error of each another. While only a small fraction, 1% of the 240,000+ total observations are extreme or influential enough to render these standard errors unreliable. Genuine cross-validation using five test observations held out at each station yields no clear preference between the models.

Next, we compare goodness of fit between the two models. Figure 4 shows the distribution (across stations) of R-squared values, along with a map of these values, indicating spatial patterns in model fitness for the (5,7) model. For our purposes, we define R-squared as

$1 - SSE/SST$, where SSE is calculated as the posterior mean of the sum of squared residuals for the station's time series, and SST is the usual total sum of squared deviations from the station's mean log-NO_3 (null model). It measures the reduction in variability attributable to the model in (1). Because this is not the standard coefficient of determination from least-squares regression, it is possible for the "reduction" to exceed the null variability (causing R-squared to be negative).

Most stations report a modest fit, though many stations with low R-squared values exhibit discernible trends and cycles; see the Supplementary Materials for examples of fits to time series. We note that the high R-squared values are scattered throughout the study area, although two pockets of consistently strong annual cycles along the upper west coast (Pays de la Loire) and central-eastern (Burgundy) regions are evident on the map (right panel of Fig. 4). Poor fits include extreme outliers and series whose dynamics undergo changes during the observation window. The 418 stations with negative R-squared values are scattered across France with no clear spatial pattern (not shown). Corresponding plots from the (7,11) model (not shown) are nearly indistinguishable from the (5,7) fit.

The distribution, across stations, of the percentage change in R-squared between the baseline (5,7) model and the (7,11) alternative provides a direct comparison between model fits. The median improvement offered by the (7,11) fit is less than 1%, with quartiles at -1.1% and $+1.5\%$, and 87% of the stations show less than 10% change.

Posterior mean point estimates of cluster-representative trends and amplitudes split in a logical fashion progressing from $K_\beta = 5$ to $K_\beta = 7$ and $K_\alpha = 7$ to $K_\alpha = 11$, as each estimated value has clear descendants when the number of clusters increases. Amplitude values appear not to exhibit truly discrete (clustering) behavior, but rather a refinement of continuously distributed values. The extreme clusters exhibit estimated trends that are somewhat attenuated for several stations, which we consider a feature of the model. The extreme trends and large amplitudes are often genuine, though several stations contain unreliable data, causing them to erroneously join the extreme clusters.

Another approach to assessing the appropriate number of clusters is to examine posterior uncertainty in how the 4450 stations are partitioned, and the degree to which these partitions coincide with model-specified clusters. Using MCMC samples of cluster membership in each of the three clustering characteristics (trend, amplitude, phase), we found optimal partitions by minimizing the variation-of-information loss with the SALSO algorithm of Dahl et al. (2021, 2022).

Co-clustering tendencies among the zero and low positive trends suggest that these could be combined into a single cluster, leaving only four effective trend clusters. The estimated partition of stations on amplitudes maintains excellent agreement with model-specified clusters in both models. Co-clustering patterns among phases lead to five main phase clusters that are associated with peak NO_3 occurring primarily in late fall through winter and early spring. Overall, posterior analysis of partitions appears to favor using fewer clusters (with exception of amplitude).

The ability to distinguish between groups with the most extreme trends/amplitudes is one reason to use the (7,11) model. However, adding trend and amplitude clusters notably fails to substantially improve the overall fit and yields materially equivalent inferences. Consistent with our objective to cluster trend and cyclical behaviors of rivers in an interpretable way,

we select the (5,7) model for all inferences reported hereafter. Corresponding results from the alternate model are consistent with those reported here.

3.2. CLUSTERING RESULTS

The fit with $K_\beta = 5$, $K_\alpha = 7$, and $K_\phi = 12$ separates stations into groups that are generally identified by the model. Stations tend to co-cluster across zero and positive trends, while amplitude clusters are clearly separated, and membership often straddles across adjacent phase clusters. In this section, we focus on station-specific estimates of the clustered time-series characteristics and defer analysis of the posterior distribution over partitions of stations to the Supplementary Materials. How well the model fits the data is positively correlated with posterior classification probabilities for amplitude and phase, as expected for any discretization of a continuously varying quantity.

Figure 5 shows posterior mean point estimates of station-specific intercepts, trends and amplitudes (averaged over cluster membership), and maximum a posteriori (MAP) point estimates of phase, across the study area. Point opacity indicates R-squared values. Several distinct river networks are discernible on the plots as veins with similar color (e.g., along the northern tip of the northeastern border, in the Hauts-de-France region), which are identified solely by information in their NO_3 series, geographic, and catchment characteristics.

Intercepts are the only parameters in Fig. 5 that are not clustered by the model. They represent the mean level (i.e., the midpoint between peak and trough) of log-NO_3 at the beginning of the observation window: January 1, 2010. The spatial distribution of intercepts positively correlates with arable land, as higher levels are generally seen in the north and northwest parts of France.

Multi-year trends appear to be absent or weak at most locations, as seen in the upper-right panel of Fig. 5. Although the trends are given for log-NO_3, estimates have small magnitudes, easing interpretation. For example, a trend coefficient of 0.05 roughly corresponds with a 5% annual increase in mean NO_3 concentration (in mg/L). Approximately 12% of stations have posterior probability (PP) greater than 0.95 of clustering outside the central cluster, for which the trend value is fixed at zero, and the lowest positive cluster; 11.3% and 0.4% of stations have high PP of belonging to other clusters with negative (decreasing NO_3) and high positive trends, respectively. Note that the MCMC chains of trends and associated cluster membership have low effective sample sizes and that the low positive trend is essentially indistinguishable from zero. While trends appear to be sustained for many of these stations, they are based on a short window of at most seven years, which makes them somewhat sensitive to decadal-scale climatic variability (Mellander et al. 2018; Dupas et al. 2016).

Prominent clusters of negative trends appear in the southwestern (Occitanie) and southeastern (Provence-Alpes-Côte d'Azur) regions. Weak negative trends are also common along the eastern section of the country. Positive trends appear primarily along the central-western (Pays de la Loire) and northern regions. While several of the sites with extreme trend estimates contain unreliable data, many represent genuine trends over some sub-interval of the seven-year window.

Clusters with high amplitudes (of log-NO_3), indicating large fluctuations, concentrate primarily in the southwestern (Occitanie) and central-western (Pays de la Loire) regions,

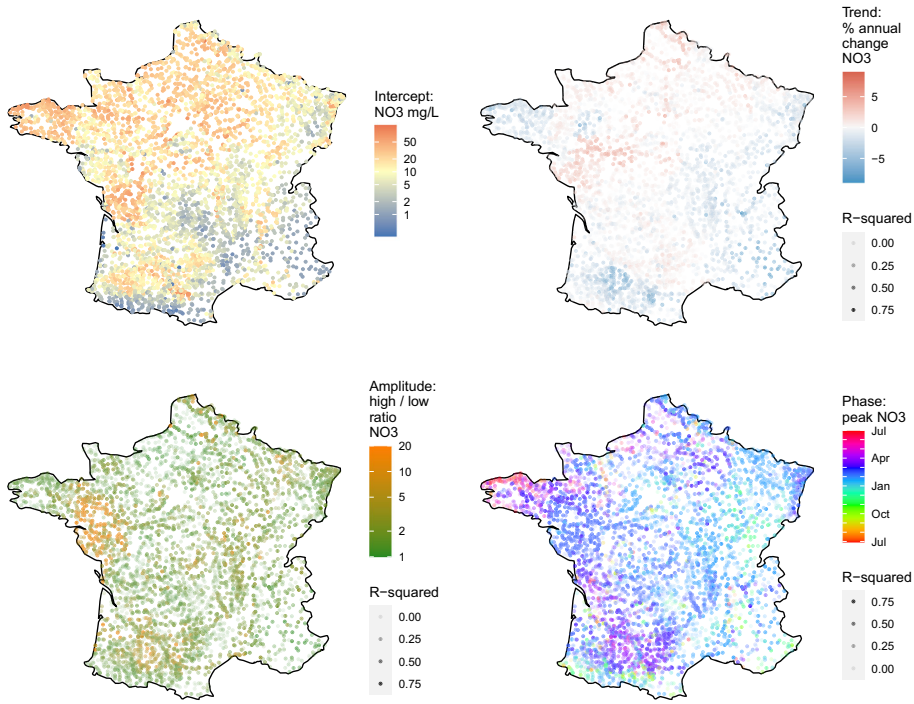

Figure 5. Model estimates representing the intercept (top left, with mean ≈ 10 mg/L), trend (top right), amplitude (bottom left), and phase (bottom right), for each station. Reported parameter values are transformed, for interpretability, from point estimates. Estimates are station-specific posterior means, except for phase, which gives the cluster-specific mean for the most likely cluster of each station. Point opacity reflects the value of R-squared for the corresponding series (Color figure online).

as seen in the amplitude panel of Fig. 5. Stations with high amplitudes tend also to feature the best-fitting series, with clear and prominent annual cycles. The two prominent high-amplitude clusters were previously noted for opposing multi-year trends and also have differing times of peak NO_3 concentration. The Pays de la Loire cluster straddles multiple watersheds, indicating a regional pattern likely associated with climate, surficial geology, or land use, which all extend beyond watershed boundaries.

The majority of locations see NO_3 peaks occurring between December and April (lower-right panel of Fig. 5). The western section of France tends to be dominated by cycles that peak in winter and early spring, while the inland, central section often peaks in late fall and early winter. Rivers in highland regions, along the Alps in the east, and particularly in the Pyrenees to the southwest, often peak in late fall, likely because NO_3 is diluted during snowmelt in late spring. The northern coast of the western peninsula (Brittany) uniquely clusters stations that see peak NO_3 in the early summer months. While this timing is consistent with animal husbandry and intensive row-crop production common in the region (Guillemot et al. 2020), it is distinct from the phase of similar areas in western France. The specific cause remains unknown, but the difference could be due to differences in residence time associated with

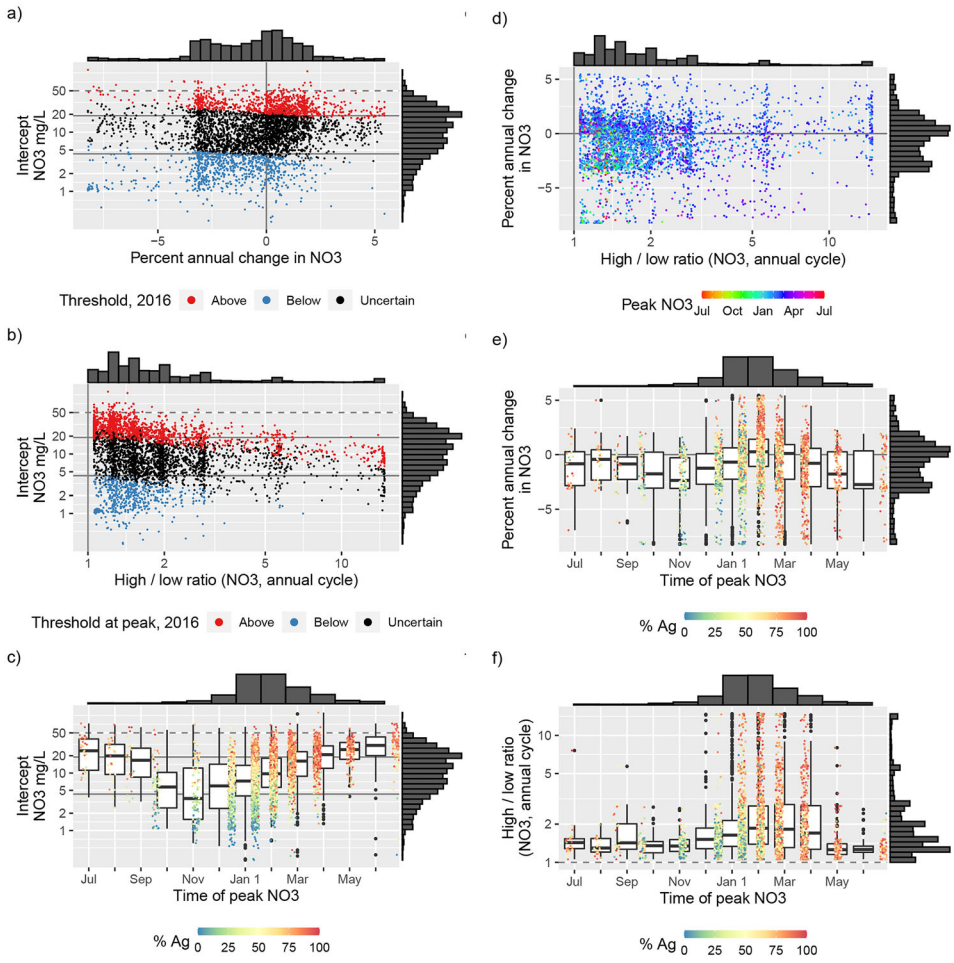

Figure 6. Scatter plots depicting relationships among parameters, by station. The left column compares the (transformed) posterior mean intercept to (transformed) point estimates of trend (**a**), amplitude (**b**), and phase (**c**). Horizontal lines on intercept axes indicate recommended NO_3 thresholds (World Health Organization, dashed; eutrophication-reduction range, solid), with accompanying hypothesis test results indicated by color. The right column compares estimates among the three clustered quantities: trend against amplitude (phase indicated by color) (**d**), trend against phase (**e**), and amplitude against phase (**f**). Box plots for each phase cluster indicate the distribution of points along the vertical axis. Histograms bordering the scatter plots show marginal distributions. Coloring in the remaining plots give percentage of land dedicated to agriculture in each catchment (Color figure online).

climate, soil, and geology, which may differentially delay and convolve NO_3 dynamics (Marçais et al. 2018).

We next analyze associations among the time-series characteristics by with scatter plots of the estimates (calculated as in Fig. 5) in Fig. 6. Cluster-representative parameter values are discernible, as points tend to concentrate near these values. Trend or amplitude estimates that are not close to the representative values accompany stations whose membership straddles more than one cluster, resulting in a mixture posterior for the station's parameter.

Relationships of trend, amplitude, and phase with the intercept are shown in the three left plots in Fig. 6. Panel (a) shows a weak positive overall association between intercepts and time trends (polyserial correlation of .147 with 95% interval (.121, .172)). However, subsetting to stations below the median elevation, where most agricultural and urban activity occurs (Thomas et al. 2016), eliminates most points in the lower-left section of the plot, weakening the correlation to 0.015 (−0.021, 0.052). This agrees with regional studies of lower-elevation catchments that found little or no relationship between initial NO_3 and trend following reduction in nutrient loading (Abbott et al. 2018b; Dupas et al. 2018).

To provide context about whether the decreases in NO_3 are sufficient to attain regulatory targets rapidly (i.e., within six years), red points in panel (a) indicate stations where median NO_3 (intercept and trend only) in 2016 exceeds the boundary of 19 mg/L (higher solid horizontal line), which is an upper limit of when reductions of eutrophication may begin (Dodds et al. 1998; Perrot et al. 2014; Abbott et al. 2018b), with posterior probability above 0.9. Blue points indicate stations below 4.3 mg/L, a lower limit for eutrophication reduction, with 0.9 posterior probability. Coloring in panel (b) indicates whether annual peaks of NO_3 are outside these thresholds with high posterior probability.

The negative relationship between intercept and amplitude (polyserial correlation −.158 (−.169, −.149) that becomes stronger when subset to low-elevation stations) is consistent with nutrient saturation syndrome (Earl et al. 2006), where overloaded watersheds show little seasonal variation, and lower nutrient watersheds show annual and diel fluctuations associated with source limitation and biological uptake (e.g., stream metabolism; Wollheim et al. 2018). Alternatively or additionally, agricultural regions with low topographical slope could have more diversity in hydrological transport time from soil to stream (Sebilo et al. 2013; Marçais et al. 2018), and greater vertical nutrient gradients (Ebeling et al. 2021) resulting in destructive interference of temporally discrete nutrient delivery signals (e.g., fertilizer application or crop harvesting) and attenuated nutrient variability in the river (Abbott et al. 2018a). The negative correlation among intercept and amplitude is particularly evident along northwest Atlantic coast (see Fig. 5), where amplitudes in Brittany might more closely resemble those further down the coast except for consistently high NO_3. As noted, patterns in this region could also result from geological variation, wherein distinct formations create hydrological compartments with distinct flow and denitrification patterns (Ben Maamar et al. 2015; Bochet et al. 2020; Kolbe et al. 2019).

There is a nonlinear relationship between intercept and phase. Sites with lowest NO_3 intercepts are much more likely to experience peak concentration in the winter, while some high-intercept sites have peaks throughout the year (Fig. 6c). Some of these sites that peak in winter also show the highest seasonal amplitude (Fig. 6f). This is again consistent with nutrient saturation syndrome, though it could also be associated with negative correlations between agricultural activity (indicated with color in panels (c), (e), and (f)) and topographical slope (i.e., it is less common to find extensive nutrient loading in mountainous regions with snowmelt-dominated NO_3 seasonality). Because agricultural coverage and type are distributed based on local climate conditions, these types of collinearities are common at regional to continental scales (Thomas et al. 2016).

The three plots on the right side of Fig. 6 elucidate co-clustering behaviors among trend, amplitude, and phase. Little-to-no global association exists between amplitude and trend

(polychoric correlation −.007 (−.038, .024), and at low-elevation −0.005 (−.051, .040)). Note, however, that locations with positive trends are over-represented in the high-amplitude group, potentially indicating the pollution of low-nutrient sites with previously high seasonal fluctuations in concentration. More pronounced relationships are evident among trend phase and amplitude phase. In both cases, extremes are more common at locations that peak in the winter and early spring.

While the vast majority of stations cluster in groups with negligible trends and low amplitudes, we are interested in groupings with extreme parameter values, which could indicate successes or failures in recent management. Three such clusters correspond to areas already mentioned: the Occitanie (southwestern) region, with negative trends, high amplitudes, and peaks in late spring; the Brittany (northwestern peninsula) region, with weak negative trends, typical amplitudes, and summer peaks along the coast; and the Pays de la Loire (central-western) region, with some weak positive trends, high amplitudes, and winter peaks.

3.3. COVARIATE EFFECTS ON CLUSTER ASSIGNMENT

We next investigate whether the static, site-specific catchment covariates described in Table 1 systematically correlate with the patterns observed in the cluster analysis. Modeling these relationships through the probability of cluster membership (as described in Sect. 2) establishes rank-type correlations, accommodating potentially nonlinear, monotonic effects. Analysis using station-specific posterior means of the clustering parameters suggests the possibility of nonlinear relationships that appear monotonic.

Three percent of sites (153 total) were missing one or two covariate values among bedrock, IDRP, and runoff. With few missing values, we opted to impute from complete cases using random forests trained on the other covariates. The tuning parameters for the random forest imputers were determined using 10-fold cross validation.

All non-binary covariates were centered and scaled in $\boldsymbol{x}_r$. Some correlation exists among these covariates, but not enough for concern (see the Supplementary Materials). Station elevation and the percent of catchment covered by water (% water body) were log-transformed prior to scaling. Because the land-use percentages are compositional, they were preprocessed via transformation along the first three principal components. The three estimated effects were then transformed back for each MCMC draw to obtain four estimated coefficients, interpretable as effects from the original land-use variables.

Table 2 reports point estimates and credible intervals for coefficients on each of the trend, amplitude, and phase clustering regressions. We define an effect as significant if its credible interval does not cover zero. Effects from departments (geographical administrative divisions of France) are omitted from the table. However, many such effects are significant (see the Supplementary Materials for maps with the estimates), as the spatial patterns in parameter estimates often roughly coincide with these geopolitical boundaries. The left panel of Fig. 7 demonstrates how department provides a useful instrument for capturing spatial information. It is also clear that beyond these block effects, heterogeneity also exists within departments, and clusters often span across multiple departments.

Table 1. Description of static covariate measurements for each monitoring site. With exception of site-specific elevation, values are aggregated to the level of the catchment associated with a station

Category	Variables (units)	Description	Source
Geopolitical	Department	The second level of administrative division in France, after Region, totaling 96 units	https://www.data.gouv.fr/fr/datasets/contours-des-departements-francais-issus-d-openstreetmap/
Elevation	(m)	Europe terrain data produced using Copernicus data; funded by the European Union - EU-DEM layers	https://land.copernicus.eu/imagery-in-situ/eu-dem/eu-dem-v1.1/view
Lithology	% Sedimentary % Igneous	Bedrock classification	http://www.geocatalogue.fr/Detail.do?id=6388
Hydrology	IDPR	Groundwater connectivity, reported through the Network Development and Persistence Index	http://www.geocatalogue.fr/Detail.do?id=13039
	Annual specific runoff (mm/year)		http://geowww.agrocampus-ouest.fr/geonetwork/apps/georchestra/?uuid=518b3e0a-ee55-40cb-a3ed-da00e60505aa
Hydrologic typography	% Water body % Wetland	Catchment water cover classification	http://geowww.agrocampus-ouest.fr/geonetwork/apps/georchestra/?uuid=518b3e0a-ee55-40cb-a3ed-da00e60505aa
CORINE land cover	% Agriculture % forest % urban % other	Catchment land composition, published 2012	https://www.geoportail.gouv.fr/donnees/corine-land-cover-2012
Biogeographical region		Classification designated by the European Environment Agency	https://www.eea.europa.eu/data-and-maps/data/biogeographical-regions-europe-3

Table 2. Posterior median estimates and 95% credible intervals for coefficients relating station-specific covariates to cluster membership probability, for each of trend, amplitude, and both directions of phase

	Trend ($\hat{\theta}_\beta$)		Amplitude ($\hat{\theta}_\alpha$)	
Intercept	1.83	(1.12, 2.57)	0.15	(−0.34, 0.64)
ln elevation	**−0.18**	(−0.30, −0.07)	**−0.12**	(−0.19, −0.05)
% sedimentary bedrock	0.02	(−0.05, 0.10)	0.03	(−0.02, 0.07)
IDPR	**−0.27**	(−0.36, −0.19)	*0.17*	(0.12, 0.22)
Annual specific runoff	*0.15*	(0.05, 0.25)	**−0.14**	(−0.20, −0.08)
ln % water body	−0.04	(−0.11, 0.03)	*0.24*	(0.20, 0.27)
ln % wetland	0.05	(−0.04, 0.14)	*0.18*	(0.12, 0.23)
% agriculture	*0.05*	(0.00, 0.10)	*0.03*	(0.00, 0.06)
% forest	**−0.08**	(−0.14, −0.03)	**−0.04**	(−0.06, −0.01)
% urban	0.05	(−0.02, 0.12)	0.02	(−0.02, 0.06)
% other land use	0.04	(−0.04, 0.11)	−0.02	(−0.07, 0.03)
Bioregion: Atlantic	**−0.87**	(−1.38, −0.37)	−0.21	(−0.55, 0.13)
Bioregion: continental	**−0.82**	(−1.34, −0.32)	−0.30	(−0.65, 0.05)
Bioregion: Mediterranean	−0.52	(−1.09, 0.06)	**−0.42**	(−0.80, −0.02)
	Phase (D_1; $\hat{\theta}_{\phi 1}$)		Phase (D_2; $\hat{\theta}_{\phi 2}$)	
Intercept	2.98	(2.06, 3.96)	1.79	(1.05, 2.53)
ln elevation	−0.08	(−0.20, 0.05)	−0.02	(−0.14, 0.09)
% sedimentary bedrock	*0.18*	(0.10, 0.26)	*0.14*	(0.06, 0.22)
IDPR	−0.06	(−0.15, 0.03)	**−0.09**	(−0.18, 0.00)
Annual specific runoff	*0.35*	(0.22, 0.48)	**−0.16**	(−0.26, 0.05)
ln % water body	*0.15*	(0.08, 0.24)	*0.36*	(0.29, 0.44)
ln % wetland	*0.27*	(0.18, 0.37)	0.02	(−0.07, 0.11)
% agriculture	**−0.19**	(−0.25, −0.14)	*0.20*	(0.15, 0.25)
% forest	*0.06*	(0.00, 0.13)	**−0.24**	(−0.29, −0.18)
% urban	*0.21*	(0.12, 0.31)	0.05	(−0.02, 0.13)
% other land use	*0.32*	(0.11, 0.59)	0.00	(−0.08, 0.10)
Bioregion: Atlantic	−0.62	(−1.34, 0.06)	**−0.82**	(−1.33, 0.31)
Bioregion: continental	0.02	(−0.72, 0.74)	**−0.92**	(−1.49, −0.34)
Bioregion: Mediterranean	−0.49	(−1.30, 0.30)	**−1.18**	(−1.86, −0.53)

Movement in the positive D_1 direction (x-axis) pushes a phase (peak NO_3) toward January 1. Movement in the positive D_2 direction (y-axis) pushes a phase toward April 1. Significant effects are bold and italic

Although coarser, biogeographical regions provide alternate, complementary spatial classifications to the less ecologically defined department boundaries. Biogeographical region classification appears to influence both NO_3 trends and phases. Negative coefficients for the D_2 (vertical) axis on phase suggest that non-alpine region indicators push peaks toward the autumnal equinox. This is somewhat surprising, given that Fig. 7 suggests high-elevation stations tend to peak in fall. These maps, however, view the relationships marginally and do not control for the other, more geographically precise covariates.

When we view clustering of time trends as the response, significant effects appear consistent with prevailing thought that lowland rivers (negative elevation effect) surrounded by agricultural activity (positive agriculture land-use effect) are among the most impacted by excess nutrients. These lowland rivers also tend to have low river density, as measured by IDPR (negative effect). The agriculture-trend relationship is not universal, as high intercepts and negative trends on the western peninsula (Brittany) indicate rapid recovery. On the other

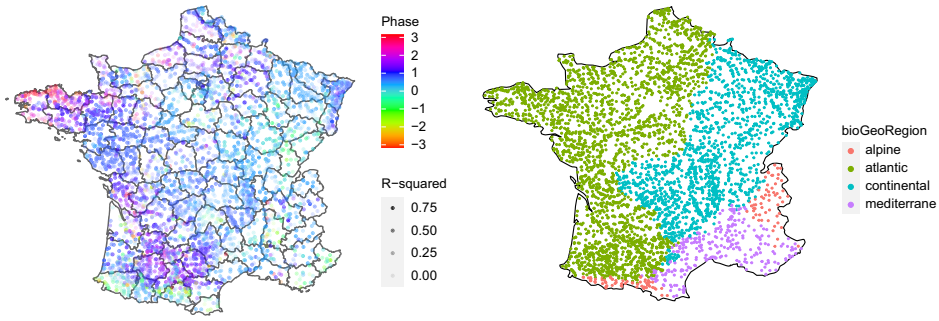

Figure 7. Estimated values of phase with department boundaries for reference (left), and biogeographical region classification (right), by station. Phase values range from $-\pi$ to $+\pi$, with 0 indicating January 1, and positive values progressing through spring to early summer (Color figure online).

hand, low-lying rivers near the northern border with Belgium exhibit high NO_3 levels and positive trends, indicating continued overloading or nutrient legacy time lags (Ebeling et al. 2021; Newcomer et al. 2021; Van Meter et al. 2018).

As with trends, amplitudes tend to be greater at lower elevations. They also exhibit positive association with the proportion of the associated catchment that is covered in water. Significant effects on amplitude associated with IDPR (positive) and runoff (negative) are reversed from their effects on trend. Land use has a less pronounced effect on amplitude than on trend.

Phase is more challenging, both from modeling and from hydrological standpoints. Interpretation of the effects coincides with the directions in Fig. 3, where positive effects along the x-axis (labeled D_1) push the phase (peak) toward the winter solstice, and positive effects along the y-axis push the phase toward spring equinox. Covariate effects on phase are likely modulated by river flow volume, which has direct relationships with NO_3 concentration and export (Frei et al. 2020; Ebeling et al. 2021), but is not included in this analysis.

The significant effects of agriculture on both axes, negative in x and positive in y, mean that catchments that are more dedicated to agriculture tend to peak in late spring. This may coincide with fertilizer application prior to maximum plant demand, which occurs in spring, although its effect on NO_3 concentration can be delayed in some regions (Aquilina et al. 2012; Dupas et al. 2020). The runoff effects oppose those of % agriculture for both axes, nudging phase probabilities for high-runoff catchments to favor late fall, potentially associated with the release of nutrients following harvesting in these catchments with greater hydrological connectivity between land and water (Zarnetske et al. 2018; Covino 2017). Other notable effects include bedrock composition and percent water cover, which appear to help account for the unique phase-clustering behavior on the northern coast of Brittany.

Because runoff types vary by geography, it would be appealing to estimate an interaction effect with biogeographical region and runoff. However, many of the predictors co-vary spatially, which limits diversity in the predictor space and introduces multicollinearity when

considering interactions. Thus, we retain additive models to preserve stability and interpretability of covariate effects.

4. CONCLUSION

We tested a unified, probabilistic framework to summarize NO_3 time series from thousands of locations across France. By clustering regression coefficients for each monitoring station, we provided a data-driven discretization of key time-series characteristics to (1) facilitate interpretation and (2) capture spatial correlation and borrow strength within river networks when explicit network information is unavailable. The methodology distilled salient information from a conceptually simple model, extracting general patterns at regional and national levels. This approach provides insight into hydrochemical behavior of individual watersheds, including multi-year trends. Furthermore, we have related the key characteristics of NO_3 concentration to several covariates that describe the catchment for each station. These regressions are applied to an ordered clustering, allowing for monotonic, nonlinear relationships. This simple and robust approach could aid in the interpretation and application of large-scale hydrochemical data by characterizing nutrient regimes and providing several metrics for gauging management successes and failures.

This analysis identified a handful of regional, distinct regimes of NO_3 dynamics including in the Occitanie (southwest; characterized by negative trends from high initial levels, and high amplitudes), Brittany (northwestern peninsula; characterized by elevated but decreasing NO_3, with low amplitudes and unique timing of peak concentration), and Pays de la Loire (central-western; characterized by elevated and increasing NO_3 with high amplitudes) regions. Although land use and hydrologic connectivity consistently correlate with patterns in trends and seasonality and help account for heterogeneity at the national level, finer-detailed geologic and water-flow covariates would be necessary to untangle complex seasonal regimes, such as those found in Brittany. Tests revealed clear negative trends in approximately 11% of stations and high uncertainty associated with positive trends.

Throughout our analysis, we have used a fixed number of clusters to describe the temporal behavior of NO_3 at spatially distinct stations. While we took great care to estimate an appropriate number of fixed clusters (see Sect. 3.1 and related Supplementary Materials) for these particular data, we acknowledge alternative approaches that could treat the number of clusters as an additional parameter to be estimated from the data. One such example is the profile regression approaches of Liverani et al. (2015) and Liverani et al. (2016) which use a latent clustering on response and explanatory variables jointly to infer their relationship.

The emergence of clusters with distinct temporal regimes, which correspond with regional agricultural practices (Poisvert et al. 2017), demonstrates how this type of national-scale analysis can aid regulators and policymakers in their efforts to assess and eventually improve water quality. Because the nutrient retention capacity of different catchments is influenced by both inherent ecological parameters and human management practices, classifying temporal trends can inform both the upper limit of sustainable nutrient loading and the effectiveness of restoration activities.

While it is highly desirable to encapsulate nutrient dynamics at the national and local levels simultaneously in reasonable computing time, the chosen model trades precision at the local level by restricting station-specific parameters to cluster-representative values, to borrow strength and achieve hard clustering. The observation model for each time series is indeed simplistic, capturing only a linear trend and an annual cycle, with autocorrelated error. We note three prominent issues.

First, exploratory work with longer time series confirms the existence of nonlinear long-term trends. However, most series in the present study cover four to seven years. Admitting more flexible trends improves local fit at the expense of additional computation and complicating interpretation. For these reasons, we restrict attention to linear trends, which are straightforward to cluster.

Second, spectral analysis of the raw time series revealed sub-annual signals, contributing primarily as deformations to the dominant annual cycle. Inclusion of sub-annual cycles would allow more detailed exploration and more precise modeling at the observation level, again at the cost of complicating interpretation. A related issue is that signals with a period of one year are not necessarily captured by a single sine wave. For example, several series exhibit brief impulse-type behavior wherein otherwise stable nitrate levels abruptly spike at the same time each year. These behaviors may point to anthropogenic interference, such as intermittent damming, or could be related to flashy hydrological behaviors in steep alpine and Mediterranean catchments.

Third, Gaussian errors are inadequate for approximately one-third of the stations. This is often due to outlier NO_3 measurements. The cyclical nature of the data also makes it difficult to distinguish outliers from evolving amplitudes, which in our model can also be confounded with trends.

To accommodate such departures from the standard model and to further classify stations, one appealing direction could be to extend the mixture model to allow clustering over alternate functional forms and error structures. As an example, one could specify a mixture component that replaces the trigonometric basis with a wavelet basis. Another mixture specification could replace Gaussian errors with those from a longer-tailed distribution, yielding more robust inferences and allowing the model to flag stations with extreme outlying measurements.

Additional mixture components assign static features to each time series. However, several rivers exhibit dynamic shifts in key characteristics. For example, some trends are not constant for the entire seven-year period, and some rivers experience abrupt changes in the amplitude of the annual nitrate signal. Capturing such shifts is beyond the scope of this analysis, particularly due to concerns about computational feasibility and cluster interpretability. Nevertheless, scalable inference for clustered dynamic models remains an appealing goal.

From an ecological perspective, one major limiting factor in this analysis is the lack of flow time series at each station. Water flow is known to influence nutrient concentrations in rivers, and knowledge of the flow volume at a particular location aids with interpretation of concentration dynamics. Expanding river discharge monitoring and modeling would enhance our ability to identify key export periods and assess management effectiveness accordingly.

SUPPLEMENTARY MATERIALS

APPENDIX

Computing strategy, including complete conditional distributions for the algorithm; MCMC implementation and diagnostics; further details on model selection and fit; partition analysis; and additional details on the static covariates. (PDF document)
R Code and Data: River-NO_3clust folder (.zip archive). Also available in GitHub repository https://github.com/mheiner/River-NO3clust.git

ACKNOWLEDGEMENTS

The authors gratefully acknowledge helpful comments from three anonymous referees and an associate editor, as well as helpful input from Candace Berrett. French water chemistry time series were provided by Naïades, extraction date 24 Nov. 2018, url: http://www.naiades.eaufrance.fr/france-entiere#/. France DEM terrain data used for elevation were provided by Copernicus, funded by the European Union - EU-DEM layers. B.W.A. was supported by the US National Science Foundation Grant No. EAR 2011439.

[Received March 2022. Revised June 2022. Accepted July 2022. Published Online September 2022.]

REFERENCES

Abbott BW, Gruau G, Zarnetske JP, Moatar F, Barbe L, Thomas Z, Fovet O, Kolbe T, Gu S, Pierson-Wickmann A-C, Davy P, Pinay G (2018) Unexpected spatial stability of water chemistry in headwater stream networks. Ecol Lett 21:296–308

Abbott BW, Moatar F, Gauthier O, Fovet O, Antoine V, Ragueneau O (2018) Trends and seasonality of river nutrients in agricultural catchments: 18 years of weekly citizen science in France. Sci Total Environ 624:845–858

Albert JH, Chib S (1993) Bayesian analysis of binary and polychotomous response data. J Am Stat Assoc 88:669–679

Álvarez-Cabria M, Barquín J, Peñas FJ (2016) Modelling the spatial and seasonal variability of water quality for entire river networks: Relationships with natural and anthropogenic factors. Sci Total Environ 545:152–162

Aquilina L, Vergnaud-Ayraud V, Labasque T, Bour O, Molénat J, Ruiz L, de Montety V, De Ridder J, Roques C, Longuevergne L (2012) Nitrate dynamics in agricultural catchments deduced from groundwater dating and long-term nitrate monitoring in surface-and groundwaters. Sci Total Environ 435:167–178

Ascott MJ, Gooddy DC, Fenton O, Vero S, Ward RS, Basu NB, Worrall F, Van Meter K, Surridge BW (2021) The need to integrate legacy nitrogen storage dynamics and time lags into policy and practice. Sci Total Environ 781:146698

Banerjee S, Carlin BP, Gelfand AE (2014) Hierarchical modeling and analysis for spatial data. CRC Press

Ben Maamar S, Aquilina L, Quaiser A, Pauwels H, Michon-Coudouel S, Vergnaud-Ayraud V, Labasque T, Roques C, Abbott BW, Dufresne A (2015) Groundwater isolation governs chemistry and microbial community structure along hydrologic flowpaths. Front Microbiol 6:1457

Berrett C, Calder CA (2012) Data augmentation strategies for the Bayesian spatial probit regression model. Comput Stat Data Anal 56:478–490

Berrett C, Calder CA (2016) Bayesian spatial binary classification. Spat Stat 16:72–102

Bochet O, Bethencourt L, Dufresne A, Farasin J, Pédrot M, Labasque T, Chatton E, Lavenant N, Petton C, Abbott BW (2020) Iron-oxidizer hotspots formed by intermittent oxic-anoxic fluid mixing in fractured rocks. Nat Geosci 13:149–155

Burt TP, McDonnell JJ (2015) Whither field hydrology? The need for discovery science and outrageous hydrological hypotheses. Water Resour Res 51:5919–5928

Cheng F, Van Meter K, Byrnes D, Basu N (2020) Maximizing US nitrate removal through wetland protection and restoration. Nature 588:625–630

Chiverton A, Hannaford J, Holman I, Corstanje R, Prudhomme C, Bloomfield J, Hess TM (2015) Which catchment characteristics control the temporal dependence structure of daily river flows? Hydrol Process 29:1353–1369

Conley DJ, Paerl HW, Howarth RW, Boesch DF, Seitzinger SP, Karl EKE, Lancelot C, Gene EGE (2009) Controlling eutrophication: nitrogen and phosphorus. Science 123:1014–1015

Covino T (2017) Hydrologic connectivity as a framework for understanding biogeochemical flux through watersheds and along fluvial networks. Geomorphology 277:133–144

Cowles MK (1996) Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models. Stat Comput 6:101–111

Cressie N (2015) Statistics for spatial data. John Wiley & Sons

Dahl DB, Johnson DJ, Müller P (2021) salso: search algorithms and loss functions for bayesian clustering. https://CRAN.R-project.org/package=salso. R package version 0.3.0

Dahl DB, Johnson DJ, Müller P (2022) Search Algorithms and Loss Functions for Bayesian Clustering. Journal of Computational and Graphical Statistics. https://doi.org/10.1080/10618600.2022.2069779

de Almeida R, Steiner MTA, dos Santos Coelho L, Francisco CAC, Neto PJS (2019) A case study on environmental sustainability: a study of the trophic changes in fish species as a result of the damming of rivers through clustering analysis. Comput Ind Eng 135:1239–1252

de Lavenne A, Skøien J, Cudennec C, Curie F, Moatar F (2016) Transferring measured discharge time series: large-scale comparison of top-kriging to geomorphology-based inverse modeling. Water Resour Res 52:5555–5576

Diaz RJ, Rosenberg R (2008) Spreading dead zones and consequences for marine ecosystems. Science 321:926–929

Dodds WK, Jones JR, Welch EB (1998) Suggested classification of stream trophic state: distributions of temperate stream types by chlorophyll, total nitrogen, and phosphorus. Water Res 32:1455–1462

Dupas R, Jomaa S, Musolff A, Borchardt D, Rode M (2016) Disentangling the influence of hydroclimatic patterns and agricultural management on river nitrate dynamics from sub-hourly to decadal time scales. Sci Total Environ 571:791–800

Dupas R, Minaudo C, Gruau G, Ruiz L, Gascuel-Odoux C (2018) Multidecadal trajectory of riverine nitrogen and phosphorus dynamics in rural catchments. Water Resources Research 54:5327–5340

Dupas R, Minaudo C, Abbott BW (2019) Stability of spatial patterns in water chemistry across temperate ecoregions. Environ Res Lett 14:074015

Dupas R, Ehrhardt S, Musolff A, Fovet O, Durand P (2020) Long-term nitrogen retention and transit time distribution in agricultural catchments in western France. Environ Res Lett 15:115011

Earl SR, Valett HM, Webster JR (2006) Nitrogen saturation in stream ecosystems. Ecology 87:3140–3151

Ebeling P, Kumar R, Weber M, Knoll L, Fleckenstein JH, Musolff A (2021) Archetypes and controls of riverine nutrient export across German catchments. Water Resour Res e2020WR028134

Ehrhardt S, Kumar R, Fleckenstein JH, Attinger S, Musolff A (2019) Trajectories of nitrate input and output in three nested catchments along a land use gradient. Hydrol Earth Syst Sci 23:3503–3524

Frei R, Frei KM, Kristiansen SM, Jessen S, Schullehner J, Hansen B (2020) The link between surface water and groundwater-based drinking water-strontium isotope spatial distribution patterns and their relationships to Danish sediments. Appl Geochem 121:104698

Garreta V, Monestiez P, Ver Hoef JM (2010) Spatial modelling and prediction on river networks: Up model, down model or hybrid? Environmetrics 21:439–456

Guillemot S, Fovet O, Gascuel-Odoux C, Gruau G, Casquin A, Curie F, Minaudo C, Strohmenger L, Moatar F (2020) Spatio-temporal controls of C-N-P dynamics across headwater catchments of a temperate agricultural region from public data analysis. In: Hydrology and earth system sciences discussions, pp 1–31. In press

Hannah DM, Abbott BW, Khamis K, Kelleher C, Lynch I, Krause S, Ward AS (2022) Illuminating the 'invisible water crisis' to address global water pollution challenges. Hydrol Process 36:e14525

Hartmann A, Mudarra M, Andreo B, Marín A, Wagener T, Lange J (2014) Modeling spatiotemporal impacts of hydroclimatic extremes on groundwater recharge at a Mediterranean karst aquifer. Water Resour Res 50:6507–6521

Higgs MD, Hoeting JA (2010) A clipped latent variable model for spatially correlated ordered categorical data. Comput Stat Data Anal 54:1999–2011

Isaak DJ, Peterson EE, Ver Hoef JM, Wenger SJ, Falke JA, Torgersen CE, Sowder C, Steel EA, Fortin M-J, Jordan CE (2014) Applications of spatial statistical network models to stream data. Wiley Interdiscip Rev Water 1:277–294

Jasra A, Holmes CC, Stephens DA (2005) Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. Stat Sci 20:50–67

Jiang L, Nielsen K, Dinardo S, Andersen OB, Bauer-Gottwein P (2020) Evaluation of Sentinel-3 SRAL SAR altimetry over Chinese rivers. Remote Sens Environ 237:111546

Kim SE, Seo IW (2015) Artificial neural network ensemble modeling with conjunctive data clustering for water quality prediction in rivers. J Hydro-Environ Res 9:325–339

Kirchner JW, Neal C (2013) Universal fractal scaling in stream chemistry and its implications for solute transport and water quality trend detection. Proc Natl Acad Sci 110:12213–12218

Kolbe T, de Dreuzy J-R, Abbott BW, Aquilina L, Babey T, Green CT, Fleckenstein JH, Labasque T, Laverman AM, Marçais J (2019) Stratification of reactivity determines nitrate removal in groundwater. Proc Natl Acad Sci 116:2494–2499

Kunkel D, Peruggia M (2020) Anchored Bayesian Gaussian mixture models. Electr J Stat 14:3869–3913

Le Moal M, Gascuel-Odoux C, Ménesguen A, Souchon Y, Étrillard C, Levain A, Moatar F, Pannard A, Souchu P, Lefebvre A (2019) Eutrophication: A new wine in an old bottle? Sci Total Environ 651:1–11

Liverani S, Hastie DI, Azizi L, Papathomas M, Richardson S (2015) PReMiuM: an R package for profile regression mixture models using Dirichlet processes. J Stat Softw 64:1–30

Liverani S, Lavigne A, Blangiardo M (2016) Modelling collinear and spatially correlated data. Spat Spatio-temporal Epidemiol 18:63–73

Lloyd C, Freer J, Collins A, Johnes P, Jones J (2014) Methods for detecting change in hydrochemical time series in response to targeted pollutant mitigation in river catchments. J Hydrol 514:297–312

Marçais J, Gauvain A, Labasque T, Abbott BW, Pinay G, Aquilina L, Chabaux F, Viville D, de Dreuzy J-R (2018) Dating groundwater with dissolved silica and CFC concentrations in crystalline aquifers. Sci Total Environ 636:260–272

Mellander P-E, Jordan P, Bechmann M, Fovet O, Shore MM, McDonald NT, Gascuel-Odoux C (2018) Integrated climate-chemical indicators of diffuse pollution from land to water. Sci Rep 8:1–10

Messer TL, Birgand F, Burchell MR (2019) Diel fluctuations of high level nitrate and dissolved organic carbon concentrations in constructed wetland mesocosms. Ecol Eng 133:76–87

Minaudo C, Dupas R, Gascuel-Odoux C, Roubeix V, Danis P-A, Moatar F (2019) Seasonal and event-based concentration-discharge relationships to identify catchment controls on nutrient export regimes. Adv Water Resour 131:103379

Moatar F, Abbott BW, Minaudo C, Curie F, Pinay G (2017) Elemental properties, hydrology, and biology interact to shape concentration-discharge curves for carbon, nutrients, sediment, and major ions. Water Resour Res 53:1270–1287

Moatar F, Floury M, Gold AJ, Meybeck M, Renard B, Ferréol M, Chandesris A, Minaudo C, Addy K, Piffady J (2020) Stream solutes and particulates export regimes: a new framework to optimize their monitoring. Front Ecol Evol 7:516

Naïades (2018) Physicochemistry data for Whole France. Data retrieved November 2018, url: http://www.naiades.eaufrance.fr/france-entiere#/

Neal RM (2003) Slice sampling. Annal Stat 31:705–767

Newcomer ME, Bouskill NJ, Wainwright H, Maavara T, Arora B, Siirila-Woodburn ER, Dwivedi D, Williams KH, Steefel C, Hubbard SS (2021) Hysteresis patterns of watershed nitrogen retention and loss over the past 50 years in United States hydrological basins. Glob Biogeochem Cycles 35:e2020GB006777

Nuñez-Antonio G, Gutiérrez-Peña E (2005) A Bayesian analysis of directional data using the projected normal distribution. J Appl Stat 32:995–1001

Nuñez-Antonio G, Gutiérrez-Peña E, Escarela G (2011) A Bayesian regression model for circular data based on the projected normal distribution. Stat Model 11:185–201

O'Donnell D, Rushworth A, Bowman AW, Scott EM, Hallard M (2014) Flexible regression models over river networks. J R Stat Soc Ser C Appl Stat 63:47–63

Osgood RA (2017) Inadequacy of best management practices for restoring eutrophic lakes in the United States: guidance for policy and practice. Inland Waters 7:401–407

Pearse AR, McGree JM, Som NA, Leigh C, Maxwell P, Ver Hoef JM, Peterson EE (2020) SSNdesign-an R package for pseudo-Bayesian optimal and adaptive sampling designs on stream networks. PLoS ONE 15:e0238422

Perrot T, Rossi N, Ménesguen A, Dumas F (2014) Modelling green macroalgal blooms on the coasts of Brittany, France to enhance water quality management. J Mar Syst 132:38–53

Poisvert C, Curie F, Moatar F (2017) Annual agricultural N surplus in France over a 70-year period. Nutr Cycl Agroecosyst 107:63–78

Ravindran P, Ghosh SK (2011) Bayesian analysis of circular data using wrapped distributions. J Stat Theory Pract 5:547–561

Reich BJ, Ghosh SK (2019) Bayesian Statistical Methods. CRC Press, Boca Raton

Rodríguez CE, Walker SG (2014) Label switching in Bayesian mixture models: deterministic relabeling strategies. J Comput Graph Stat 23:25–45

Schaller MF, Fan Y (2009) River basins as groundwater exporters and importers: implications for water cycle and climate modeling. J Geophys Res Atmos 114

Schliep EM, Hoeting JA (2013) Multilevel latent Gaussian process model for mixed discrete and continuous multivariate response data. J Agric Biol Environ Stat 18:492–513

Sebilo M, Mayer B, Nicolardot B, Pinay G, Mariotti A (2013) Long-term fate of nitrate fertilizer in agricultural soils. Proc Natl Acad Sci 110:18185–18189

Smits AP, Ruffing CM, Royer TV, Appling AP, Griffiths NA, Bellmore R, Scheuerell MD, Harms TK, Jones JB (2019) Detecting signals of large-scale climate phenomena in discharge and nutrient loads in the Mississippi-Atchafalaya River basin. Geophys Res Lett 46:3791–3801

Sperrin M, Jaki T, Wit E (2010) Probabilistic relabelling strategies for the label switching problem in Bayesian mixture models. Stat Comput 20:357–366

Stephens M (2000) Dealing with label switching in mixture models. J R Stat Soc Ser B 62:795–809

Stoddard JL, Van Sickle J, Herlihy AT, Brahney J, Paulsen S, Peck DV, Mitchell R, Pollard AI (2016) Continental-scale increase in lake and stream phosphorus: Are oligotrophic systems disappearing in the United States? Environ Sci Technol 50:3409–3415

Thomas Z, Abbott BW, Troccaz O, Baudry J, Pinay G (2016) Proximate and ultimate controls on carbon and nutrient dynamics of small agricultural catchments. Biogeosciences 13:1863–1875

Underwood KL, Rizzo DM, Schroth AW, Dewoolkar MM (2017) Evaluating spatial variability in sediment and phosphorus concentration-discharge relationships using Bayesian inference and self-organizing maps. Water Resour Res 53:10293–10316

Van Meter KJ, Van Cappellen P, Basu NB (2018) Legacy nitrogen may prevent achievement of water quality goals in the Gulf of Mexico. Science 360:427–430

Vaughan MC, Bowden WB, Shanley JB, Vermilyea A, Sleeper R, Gold AJ, Pradhanang SM, Inamdar SP, Levia DF, Andres AS (2017) High-frequency dissolved organic carbon and nitrate measurements reveal differences in storm hysteresis and loading in relation to land cover and seasonality. Water Resour Res 53:5345–5363

Vehtari A, Gelman A, Gabry J (2017) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat Comput 27:1413–1432

Ver Hoef JM, Peterson EE (2010) A moving average approach for spatial statistical models of stream networks. J Am Stat Assoc 105:6–18

Wang F, Gelfand AE (2013) Directional data analysis under the general projected normal distribution. Stat Methodol 10:113–127

Wollheim WM, Bernal S, Burns DA, Czuba J, Driscoll C, Hansen A, Hensley R, Hosen J, Inamdar S, Kaushal S (2018) River network saturation concept: factors influencing the balance of biogeochemical supply and demand of river networks. Biogeochemistry 141:503–521

Yan D, Wang K, Qin T, Weng B, Wang H, Bi W, Li X, Li M, Lv Z, Liu F (2019) A data set of global river networks and corresponding water resources zones divisions. Sci Data 6:1–11

Zarnetske JP, Bouda M, Abbott BW, Saiers J, Raymond PA (2018) Generality of hydrologic transport limitation of watershed organic carbon flux across ecoregions of the United States. Geophys Res Lett 45:11–702

Zhang Q, Webber JS, Moyer DL, Chanat JG (2021) An approach for decomposing river water-quality trends into different flow classes. Sci Total Environ 755:143562

Zimmerman DL, Ver Hoef JM (2017) The Torgegram for fluvial variography: characterizing spatial dependence on stream networks. J Comput Graph Stat 26:253–264

Zubaidah T, Karnaningroem N, Slamet A (2018) K-means method for clustering water quality status on the rivers of Banjarmasin, Indonesia. ARPN J Eng Appl Sci 13:3692–3697