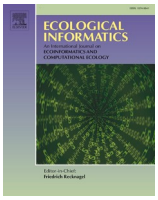
Ecological Informatics 84 (2024) 102877

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Gaussian process regression as a powerful tool for analysing time series in environmental geochemistry

Teba Gil-Díaz [a,*], Michael Trumm [b]

[a] Karlsruhe Institute of Technology, Institute of Applied Geosciences, 76131 Karlsruhe, Germany,
[b] TNG Technology Consulting GmbH, www.tngtech.com, Germany

ARTICLE INFO

Keywords:
Machine learning
Linear decomposition
Monitoring
Antimony
Garonne River

ABSTRACT

Monitoring programs require more advanced data management for the registered time series. Classical temporal series decomposition cannot fulfil current needs regarding adequate data representation, optimization of the spatial-temporal sampling resolution and predictive power. In the manuscript at hand, we will demonstrate that Gaussian process regression (GPR) models are a vital machine-learning tool to interpret temporal series, improving understanding of geochemical cycles, providing input data for geochemical models and acting as a guide for future decisions in environmental monitoring. Firstly, we explore the impacts of sampling frequency in the GPR performance for temporal series with variable lengths and sampled frequencies of water discharges. On a second approach, we present the strengths and weaknesses between classical decomposition of temporal series and GPR results for a case study: a 14-year record of water discharge, suspended particulate matter and antimony concentrations in the Garonne River. Our results suggest that (i) even short temporal series with low sampling resolution can be accurately characterized by GPR when presenting well defined seasonal patterns, and (ii) GPR provides more detailed and robust support than classical statistics to identify processes responsible for multi-scale geochemical signals. This work provides a reference for researchers, engineers, and stakeholders for more reliable monitoring, understanding, and managing aquatic ecosystems.

1. Introduction

Monitoring programs of environmental physical-chemical parameters in aquatic systems (e.g., water discharge, temperature, pH, trace element concentrations, etc.) are designed to assess water quality. These programs often monitor identified contaminated sites (i.e., dispersion of the contamination, bioavailability and potential socio-economic impact of the contaminant, effectiveness of remediation measures, system resilience, etc.) or have been designed to understand a particular system in the context of Climate Change (e.g., ocean acidification, variations in water discharges, groundwater levels, etc.). Overall, monitoring programs produce long-term temporal series of several parameters. Advanced data processing of these series is necessary and of utmost importance to allow researchers and policy makers to (i) identify adequate monitoring frequency without missing dominating local/regional processes and/or anomalies, (ii) obtain tools for early warning or extreme event identification, and (iii) predict interpolated data between sampled points, helping quantify flux transfers of biogeochemical cycles and/or future conditions/scenarios, for watershed management. The key points required to achieve these goals involve having a representative dataset, allowing to characterize properly the site, and selecting the appropriate mathematical approach for the data processing. In this work specifically, we will address points (i) and (iii).

1.1. Options for time series analysis: classical vs modern approaches

Over time, several mathematical models have been used for data treatment of temporal series. The biggest gap concerning computational power and output reliability exists between classical statistic methods (i.e., also known as traditional statistics or linear correlation and regression analyses; Khalil and Ouarda, 2009, Zhu and Piotrowski, 2020) and machine learning (ML) approaches. Classical statistics based on decomposition of temporal series (e.g., methods identifying a linear trend via Sen's slope, periodicity via a seasonal factor, and random error as noise in the series) was and is still commonly used, given its simple application and fast overview of the data (Houston, 1983, Gil-Díaz et al., 2018. Fourier analysis, Multiple Regression Analysis, ARIMA, etc., constitute examples of classical statistic approaches used for time series

* Corresponding author.
E-mail address: teba.gil-diaz@kit.edu (T. Gil-Díaz).

https://doi.org/10.1016/j.ecoinf.2024.102877
Received 8 May 2024; Received in revised form 25 October 2024; Accepted 2 November 2024
Available online 8 November 2024

decomposition and analysis (e.g., Kovač-Andrić et al., 2009). Nevertheless, classical statistics range widely in generalisability, precision, and realism, meaning that we still need to identify under which conditions classical statistics, or rather other approaches, accurately predict observational data and provide understanding of system behaviour (Nielsen, 2019; Robson, 2014).

With increasing computational power and accessibility of resources, new methods are becoming available to improve on the limitations and restraints of classical statistics. Statistical models commonly require normality or non-serial correlation for non-parametric approaches that environmental datasets do not always satisfy. Even more importantly, describing datasets by fitting the best static functions imposed by the modeller (e.g., linear regression) biases not only the understanding of the system but also the further validity and application of the model for predicting scenarios. To gain a deeper understanding of parameter correlation, environmental studies nowadays apply more frequently ML approaches, often for other applications such as clustering of variables (Di et al., 2019) and sometimes for temporal series analyses (Ding et al., 2023; Hanson et al., 2020). In any case, ML allows a wider flexibility in terms of data conditions (input/constraints) and deals in a proficient way with bigger datasets presenting complex non-linear issues, providing more reliable and powerful predictive tools (Nielsen, 2019). Deep learning or artificial neural network models are generally popular, and some works have already been published showing applications in aquatic studies, e.g., for predicting long-term groundwater levels (Wunsch et al., 2022), for detecting anomalies in multimodal time series showing both spatial and temporal dependence (Ding et al., 2023), identifying water treatment plant resilience to an organic contaminant (Gheibi et al., 2022) and even for predicting dissolved oxygen or ^{137}Cs dynamics from other hydrodynamic variables (i.e., complicated encoder-decoder approaches, Hu et al., 2024, Pelé et al., 2024). However, data availability often causes inaccuracies in the performance of neural networks (Pelé et al., 2024). In addition, neural networks rarely provide interpretable trends that shed light on the underlying data correlation due to their complex neural structure and need big amounts of data for training (Bonakdari et al., 2019)

1.2. Time series analysis with GPR models

Another ML technique, Gaussian Process Regression (GPR; Rasmussen and Williams, 2006), allows a different approach to time series. GPR is an advanced ML approach which excels at flexible non-parametric modelling, accommodating multimodal data while finding functional relationships (i.e., kernel parameters; Zhu and Piotrowski, 2020). These kernels (e.g., long-term, seasonal/decay and short-term events) describe the covariance of a priori dispersal functions over the target data points. GPR computes a likelihood function based on the training information and assumes that the joint probability distribution of the model output is Gaussian. Samples of these stochastic processes give rise to an average mean function which then can be used for interpolations between sampled points and/or predictions for future trends (Elbeltagi et al., 2021; Zhu and Piotrowski, 2020). The models also include uncertainty and hyperparameter estimates for the predictions (Camps-Valls et al., 2019), which is one of the major advantages of GPR compared to other ML methods (Zhu and Piotrowski, 2020), converting GP-based models one of the most popular tools for acquiring valid and useful predictions (Vu, 2023). In fact, out of several forecasting techniques applied to wind speed prediction, GPR has been highlighted as an improved approach capable of accounting for the impact of data stability on prediction accuracy (Hoolohan et al., 2018). Compared to neural networks, GPR models suffer less from over-fitting issues (Lin et al., 2019).

In addition to the evolving computer hardware capacities and the explosion of the machine learning community, GPR nowadays can be scaled up to further account for the computational cost growing cubic ($O(N^3)$) with the data size (i.e., a comprehensive review on scalable GPs can be found in Liu et al., 2020). Furthermore, advancements in techniques have been made such as Deep Gaussian Processes (DGPs), which constitute a combination of GPs in a layer structure, allowing to model highly complex data structures. DGPs have been employed in some studies over the recent years, directly as a 'surrogate model' or 'emulator', alternative to the original model (e.g., to support model-based Climate Change mitigation, Nortier et al., 2024) or for alternative approaches (e.g., in autoencoders, Domingues et al., 2018, Camastra et al., 2023). However, there are some disadvantages of DGPs. Mainly, interpretability as for neural networks is difficult, as deeper layers lose correlation to the actual input data. Hence, given the lack of comparison with linear regression approaches (classical statistics) and the additional high computational demand of DGPs, we refrained from applying DGPs in this study.

1.3. Objectives, novelty and impact

Despite GPR being a powerful technique, its main drawback is its computational complexity, explaining its scarce application in the scientific literature. Most applications of GPR in environmental observation systems until now are related to fields of research related to green energies (e.g., solar radiation, wind speed), where GPR is commonly used in hybrid forms with other tools such as clustering analyses and/or artificial neural networks (e.g., Hoolohan et al., 2018; Vu, 2023; Zhang et al., 2021). Other disciplines at the water-soil interface have also used GPR models for identifying the optimal number of physical parameters that describe the water footprint in the River Nile (Elbeltagi et al., 2021), or to obtain model uncertainty estimates for topsoil physical properties (Ballabio et al., 2019). Thus, until now, few studies have applied GPR models per se to characterize time series of aquatic systems.

Most of the available studies in aquatic systems have used GPR for forecasting physical parameters (e.g., water temperature, conductivity, water level, etc.) from direct or remote-sensing data (Camps-Valls et al., 2019), for predicting e.g., water level in canals (Bonakdari et al., 2019) or long-term dam deformation (Lin et al., 2019), and for forecasting river water temperature (Zhu and Piotrowski, 2020). Only one study has highlighted that, out of eight machine learning algorithms, GPR approaches show superior performance in predicting water quality scores (i.e., Sajib et al., 2024). However, to the best of our knowledge, no study has tried to provide a systematic approach and comparison on the characterization of hydro-geochemical time series with GPR, applicable to several sites in a given area, particularly for the area of study of this work. This means that current models evaluating trace element time series and predicting future concentrations for estimated long-term fluxes are often based on linear trend decomposition and temporal compartmentalization (Pougnet et al., 2019; Pougnet et al., 2022). A study on the adequacy of GPR is still missing and could offer better opportunities for more appropriate mass balance calculations in a given aquatic system, allowing to characterize more precisely the system resilience to geogenic vs anthropogenic impacts (i.e., useful for biogeochemical cycles and watershed management). In addition, GPR can also be applied for characterizing and predicting temporal series that are currently derived from more demanding fluid dynamics (e.g., Navier-Stokes equations; Ranjbar et al., 2020), simplifying the computation time and applicability of aquatic models.

Therefore, the core of this work is a proof of concept regarding the use of GPR for time series analyses in aquatic systems, applied for the first time in the context of environmental monitoring of geochemical parameters in a dynamic river system. We aim at (i) evaluating the performance of GPR as a tool for identifying temporal sampling resolution, and (ii) identifying via GPR the local/regional processes dominating geochemical reactivity and behaviour compared to classical statistics. For the latter, a comparison with classical statistics will be included to highlight and validate the advantages of GPR and the environmental implications of such approaches.

2. Material and methods

2.1. Gaussian processes

In the following we present a short introduction to Gaussian processes (GPs). For detailed information, the reader can consult Rasmussen and Williams (2006).

The target of a GP is to model a functional mapping consisting of the underlying input-output correlation (*f(x)*) as well as the noise related to, e.g., uncertainties in measurements:

$$y = f(x) + N(0, \tau^2) \tag{1}$$

Here y represents the measured target values and the noise is assumed to follow a gaussian distribution with mean zero ($N(0,\tau^2)$). GPs can hence be interpreted as an infinite set of prior probability distributions with a certain mean function *m(x)* and a covariance function *k(x, x')* with x and x' being input data to the function of interest.

$$f(x) \sim GP(m(x), k(x, x')) \tag{2}$$

To evaluate a GP on an input x^* given a set of known inputs X, a joint gaussian distribution is formed to the output y^* according to:

$$\begin{pmatrix} y \\ y^* \end{pmatrix} \sim N\left(0, \begin{bmatrix} K(X,X) + \tau^2 \mathrm{I}_n & K(X,x^*) \\ K(x^*,X) & K(x^*,x^*) + \tau^2 \end{bmatrix}\right) \tag{3}$$

with K being the covariance matrix, i.e. the kernel of the GP, defined by the mutual covariances between all data points. This leads to posterior predictive mean *E[y*]* and variance *V[y*]* given by:

$$E[y^*] = K(x^*,X)\left[K(X,X) + \tau^2 \mathrm{I}_n\right]^{-1} y \tag{4}$$

$$V[y^*] = K(x^*,x^*) - K(x^*,X)\left[K(X,X) + \tau^2 \mathrm{I}_n\right]^{-1} K(X,x^*) + \tau^2 \tag{5}$$

The set of hyperparameters Θ (lengthscales, variance, period, etc.) of the kernel functions are optimized by maximizing the marginal likelihood of the predictions $\mathbb{P} = p(y|X, \Theta)$, commonly written as:

$$log\mathbb{P} = -\frac{1}{2} y^T \left(K(X,X) + \tau^2 \mathrm{I}_n\right)^{-1} y - \frac{1}{2} log\left|K(X,X) + \tau^2 \mathrm{I}_n\right| - \frac{N}{2} log 2\pi \tag{6}$$

where N is the number of samples.

2.2. Model approach

To obtain a predictive and interpretable model that can be also transferred and compared to similar systems the employed GPR models consist of 4 parts that describe environmental phenomena. A squared exponential (SE) kernel K_{long} (e.g., timescales >1y, Eq. (7)) to describe long-term trends, a periodic kernel $K_{seasonal}$ modelling seasonality (Eq. (8)), a SE kernel K_{decay} capturing seasonal decay (Eq. (7)) and a SE kernel K_{event} (timescale <60 days, Eq. (7)) describing flood events. After successive inclusion of the kernels in the order described above, the complete kernel for the GP model is $K = K_{long} + K_{seasonal} \cdot K_{decay} + K_{event}$. This is the classical composition of a kernel, where functions are selected according to the expected, logical distribution of each component of the dataset (e.g., composition of the GPR for the Mauna Loa atmospheric CO_2 long-term record, Rasmussen and Williams, 2006). For more information on typical kernel functions used in Earth observation analyses, the reader can consult Camps-Valls et al. (2016). Overall, kernels provide information via three main parameters: the variance (i.e., related to the amplitude and spread of the values), the lengthscale (i.e., timespan within which points correlate significantly), and the periodicity (i.e., how long it takes for the time series to repeat its pattern). All models are obtained according to the same protocol, which constructs the total kernel K sequentially in the four steps. The models were implemented as described in the GPFlow library (Matthews et al., 2017).

$$K_{SE}(x,x') = \sigma^2 exp\left(-\frac{(x-x')^2}{2l^2}\right) \tag{7}$$

$$K_{seasonal}(x,x') = \sigma^2 exp\left(-\frac{2sin^2(\pi|x-x'|/p)}{l^2}\right) \tag{8}$$

2.3. Datasets and data treatment

To proof the concept, we used data from two long-term monitoring programs performed at the Garonne River (SW France), displaying both periodicity as well as short-lived flood events in several variables. The first monitoring program belongs to the National Hydrographic Bank (DIREN, http://www.hydro.eaufrance.fr/), reporting online historical records of daily river discharges (Q) at several locations, some over >50 years long. For this study, we extracted an average daily Q dataset of 21 years (Supplementary Fig. S1). The second database used in this work is a published time series (Gil-Díaz et al., 2018, Supplementary Fig. S2 and S3) that combines the river discharges (Q) with transported suspended particle matter (SPM), and antimony (Sb) concentrations in both, dissolved (Sb_d) and particulate (Sb_p) phases (i.e., more specifically, Th-normalized Sb_p to avoid particle grain size effects, as explained in Gil-Díaz et al., 2018). The latter dataset is over 14 years long with a sampling frequency of 24 days (as defined in Coynel et al., 2004). Overall, five sampling sites along the Garonne River watershed are evaluated, accounting for the Pyrenean source (i.e., Port-Sainte-Marie – PSM site), the main tributary of the Garonne (i.e., Temple – T, Boisse-Penchot – BP and a historically contaminated site in Riou Mort – RM, in the Lot River watershed), and its overall contribution to the incoming flow into the Gironde Estuary (i.e., La Réole – LR). More details about the dataset characteristics can be found in Gil-Díaz et al. (2018). For further use and comparison between sites, the data of the time series was normalized using the quartile normalization with the 0–75 % quartile range, being mapped to the (0,1] interval. Thus, this normalization is not disturbed by extreme events such as floods (i.e., outliers beyond the 75 % quartile), as they do not impact the overall dataset distribution. Appropriate normalization is a prerequisite to avoid degeneration of ML models (Passalis et al., 2020) and allows a better comparison between models for different systems/sample sites.

3. Results

3.1. Model performance

3.1.1. Sampling frequency

As a first approach, a selected time series of 21 years at the Garonne River (at site LR, Supplementary Fig. S1) was extracted from the DIREN database to test the impact of sampling frequency and amount of data on the performance of the ML algorithm. The full 21-year long time series was used to construct a model according to the protocol defined earlier considering every 32nd, 24th, 16th, 8th, 4th, and 2nd point in the dataset. No contribution of seasonal decay or long-term trends were found for this dataset and hence not mentioned in the following discussion. Results (Table 1, complete parameters shown in Supplementary Table S1) show fast convergence for the seasonal part with stable values starting from the LR_Q_32 model. Hence, our GPR models are robust when identifying the seasonal component of the series, independent of the sampling frequency (i.e., tested for up to monthly frequencies for temporal series with periodicity of one year). Due to the smoothness of the measured data, the lengthscale of the events decreases with the sampling frequency. The sampling frequency governs how certain GPR models can determine and quantify the periodicity/correlation of existing events in the river (e.g., flood events for the case of Q time series). This is in line with the improved accuracy of the GPR prediction (i.e., increased coefficient of determination, r^2, with sample frequency).

Table 1
Resulting GPR model parameters for the 21-y time series of water discharge (Q) at La Réole (LR) with varying dataset temporal frequencies. Simulated sampling frequencies of the dataset varies from every day ("_1") to every 32 days ("_32"). Parameters: Lengthscales (l), variance (σ^2), period (p), and estimated dataset noise variance (τ^2). Time scales: years (y) or days (d). Measured model accuracies: mean absolute error (MAE), root mean square deviation (RMSD), the log marginal likelihood (log(ℙ)) and the coefficient of determination between the measured and modelled data (r^2).

Model	Seasonality		Events		Noise	Model accuracy			
	σ^2	p (y)	σ^2	l (d)	τ^2	MAE	RMSD	-log(ℙ)	r^2
LR_Q_32	0.89	1.01	0.04	164	0.36	0.36	0.59	236	0.38
LR_Q_24	0.63	1.00	0.11	38	0.22	0.30	0.53	272	0.50
LR_Q_16	0.54	1.00	0.12	26	0.24	0.28	0.51	428	0.55
LR_Q_8	0.79	1.00	0.17	15	0.16	0.22	0.43	727	0.68
LR_Q_4	0.54	1.00	0.21	11	0.14	0.18	0.37	1293	0.76
LR_Q_2	0.51	1.00	0.28	4	0.07	0.10	0.22	2078	0.92
LR_Q_1	2.26	1.00	0.30	2	0.01	0.04	0.08	1097	0.99

Noteworthy, the computing time of the model scaling $O(N^3)$ with the number of samples N.

3.1.2. Dataset length

Another important feature is the length of the time series, as certain correlations only appear sporadically, which can prevent the model from learning their characteristics. To this end we selected the full 21-year long time series based on an 8-day sampling frequency (LR_Q_8_21y) and cropped it to include only the most recent 21, 14, 8, 4 and 2 years (Table 2, complete parameters shown in Supplementary Table S2). The periodicity of the seasonal event is only consistently identified for this Q timeseries from a length of the dataset longer than 4 years. Nevertheless, the accuracy of the prediction (r^2) seems less dependent on the length of the dataset. However, this cannot be generalized to all timeseries, and a case-by-case study should be done to rule out the relevance of the dataset length. Despite different time periods, the event kernel is stable in all cases within the uncertainty. Based on this analysis we determine an accuracy of ±40 days for the periodic kernel and ± 2 days for the modelled event correlation. This shows that the accuracy of the models later (i.e., database of 14 years with a 24-day sampling frequency) is sufficient to get K_{event} lengthscales with errors of several days, whereas the period is more robust in both dimensions (cf. Tables 1 and 2) and can be determined within 2 % error.

3.2. Model application for geochemical data

GPR model parameters for the second database used in this work, containing the full 14 years long geochemical time series with a sampling frequency of 24 days, for Q, SPM, Sb_d and Sb_p/Th_p concentrations at five sampling sites along the Garonne River watershed are presented in Table 3. Details on complete parameters are shown in Supplementary Table S3. A kernel is considered to have no contribution (−) when the variance (σ^2) of the kernel hyperparameters is zero after optimization. This is generally the case for the Trend (and some seasonal patterns), indicating that the long-term pattern of the datasets is not explained by a long-term correlation of the data, but rather by the other components (e.g., environmental processes dominated by the seasonal variability or sporadic events caused the observed pattern).

Overall, all sites and variables have been adequately characterized via GPR models, showing similar orders of magnitude for the model accuracies (r^2), despite the variable temporal patterns of each site (e.g., natural sites vs sites showing anthropogenic influence). In addition, model accuracies show improved values compared to the expected performance of a 24-day frequency time series, as indicated from the water discharge systematic analyses (Tables 1 and 2). The fact that GPR did not find a systematic seasonality for all variables suggests that the periodicity in the GPR is not an artifact of the method. No significant contribution of the seasonal decay was found for any dataset in this study. More comments on the implications of these GPR models are found in the discussion.

4. Discussion

4.1. GPR for monitoring geochemical systems

4.1.1. Insights to understanding environmental timeseries: linear decomposition vs GPR

Classical statistics from Gil-Díaz et al. (2018), i.e., non-parametric Mann-Kendall test for the trend, combined with Sen's slope and significance of changes with Pettitt's test, was applied to the 14-year time series for all geochemical data (Q, SPM, Sb_d and Sb_p) and five sites at the Garonne River. Supplementary Table S4 resumes the outcomes from Q and SPM time series analysis whereas Supplementary Table S5 presents the outcomes for Sb_d and Sb_p. Supplementary Fig. S4 and S5 show the graphical representations of the time series decomposition for Sb_d and Sb_p, including the model accuracies (r^2) regarding their predictive capacity (measured vs modelled results). In summary, these results reported no significant trends for Q and SPM (except BP), whereas all Sb_d and Sb_p time series had significant decreasing trends over time for all sites, with a change in slope around 2008 (except at RM, Gil-Díaz et al., 2018). In the lack of further evidence, such Sb linear trends, common among sites and independent of Q and SPM, were hypothesized to be related to potential atmospheric influence (e.g., a potential combination of anthropogenic releases with climatic oscillation patterns),

Table 2
Resulting GPR model parameters for the 21-y time series of water discharge (Q) at La Réole (LR) with varying dataset lengths. Dataset lengths vary from the complete dataset of 21 years ("_21y") to only include the last 2 years ("_2y"), at a sampling frequency of every 8 days ("_8_"). Parameters: Lengthscales (l), variance (σ^2), period (p), and estimated dataset noise variance (τ^2). Time scales: years (y) or days (d). Measured model accuracies: mean absolute error (MAE), root mean square deviation (RMSD), the log marginal likelihood (log(ℙ)), and the coefficient of determination between the measured and modelled data (r^2).

Model	Seasonality		Events		Noise	Model accuracy			
	σ^2	p (y)	σ^2	l (d)	τ^2	MAE	RMSD	-log(ℙ)	r^2
LR_Q_8_2y	1.48	0.95	0.26	15	0.31	0.27	0.52	99	0.65
LR_Q_8_4y	0.83	1.00	0.23	16	0.11	0.20	0.40	126	0.70
LR_Q_8_8y	0.71	0.99	0.14	17	0.15	0.20	0.41	256	0.65
LR_Q_8_14y	0.94	1.00	0.18	15	0.19	0.20	0.40	522	0.66
LR_Q_8_21y	0.79	1.00	0.17	15	0.16	0.22	0.43	727	0.68

Table 3
Resulting GPR model parameters for geochemical, 14-y time series in the Garonne River. Lack of contribution of a kernel are shown with a dash (–). Sites: La Réole (LR), Port-Sainte-Marie (PSM), Temple (T), Boisse-Penchot (BP) and Riou Mort (RM). Data series: water discharge (Q), suspended particulate matter (SPM), dissolved Sb (Sb_d) and Th-normalized particulate Sb (Sb_p/Th_p). Parameters: Lengthscales (l), period (p), and estimated dataset noise variance (τ^2). Time scales: years (y) or days (d). Measured model accuracies: mean absolute error (MAE), root mean square deviation (RMSD), the log marginal likelihood (log(ℙ)) and the coefficient of determination between the measured and modelled data (r^2).

Site	Variable	Trend	Seasonality		Events	Noise	Model accuracy			
		l (y)	l (y)	p (y)	l (d)	τ^2	MAE	RMSD	-log(ℙ)	r^2
	Q	–	0.61	1.01	33	0.67	0.38	0.69	337	0.73
	SPM	–	–	–	1	1.83	0.18	0.47	648	1.00
	Sb_d	–	1.49	1.00*	113	0.06	0.14	0.22	42	0.99
LR	Sb_p/Th_p	6.35	3.16	0.99	30	0.22	0.28	0.43	171	0.52
	Q	–	1.29	1.01	48	0.73	0.41	0.80	317	0.48
	SPM	–	–	–	2	6.88	0.43	1.26	729	0.98
	Sb_d	4.33	0.77	1.01	53	0.05	0.11	0.21	40	0.81
PSM	Sb_p/Th_p	3.44	–	–	23	0.14	0.17	0.26	181	0.87
	Q	–	1.44	1.01	29	0.21	0.25	0.39	203	0.75
	SPM	–	–	–	67	0.19	0.23	0.30	140	0.52
	Sb_d	5.00	1.61	0.99	90	0.11	0.18	0.32	100	0.63
T	Sb_p/Th_p	10.4	–	–	34	0.08	0.18	0.24	65	0.66
	Q	–	1.79	1.00*	34	0.18	0.24	0.36	160	0.74
	SPM	3.02	–	–	5	0.57	0.44	0.64	269	0.72
	Sb_d	14.4	0.63	0.99	44	0.22	0.23	0.42	167	0.59
BP	Sb_p/Th_p	14.8	–	–	38	0.10	0.18	0.27	99	0.73
	Q	–	3.17	1.01	9	0.74	0.33	0.58	332	0.88
	SPM	–	2.99	1.00	8	0.28	0.23	0.40	185	0.93
	Sb_d	–	1.01	1.01	58	0.38	0.37	0.57	238	0.50
RM	Sb_p/Th_p	1.09	2.56	1.00*	42	0.12	0.22	0.23	116	0.68

* Period fixed after optimization of the periodic kernel.

dominating the long-term Sb signal in the Garonne watershed. When applying the GPR model to the same time series for all geochemical data and five sites at the Garonne River, results show contrasting outcomes per variable and studied site (Table 3, Supplementary Table S3). For example, there are no systematic, common parameters for Sb_d and Sb_p/Th_p time series for all sites (except for the periodic factor at LR and RM, and the Trend at BP). However, given the better representation of the measured variables via GPR (e.g., Sb_d at LR, Fig. 1; for all sites see Supplementary Fig. S6) and the significant improvement of the model accuracy (r^2; Supplementary Fig. S4 and S5 vs S6) one can safely interpret this as the most certain outcome compared to classical statistics.

As with classical statistics, Q and SPM show no particular long-term trend over time at all sites with GPR, and the common exception for SPM at BP (i.e., variable with a significant trend in both classical statistics and GPR). Nevertheless, Sb_d and Sb_p/Th_p trends show different long-term lengthscales. Since the lengthscale is a measure of the correlation time of data points, the sites having different long-term lengthscales discourages the idea of a single anthropogenic discharge in 2008 as the sole source of the observed long-term decontamination trend (Gil-Díaz et al., 2018). In addition to the given model parameters (e.g., Table 3), the Bayesian probability averages and uncertainties of GPR models (e.g.,

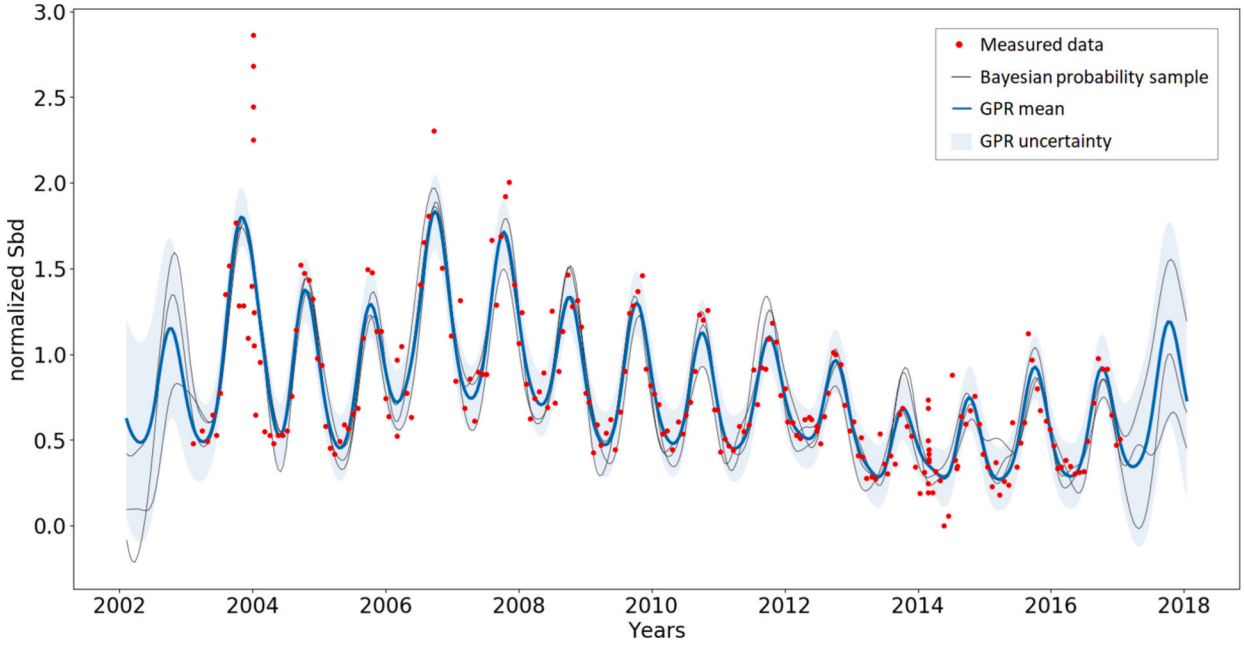


Fig. 1. Example of the GPR performance. Comparison of the 14-year temporal series of Sb_d at La Réole and the GPR visual output. The GPR uncertainty corresponds to the 2σ interval of *f(x)* including the noise (Eq. (1)). For details on GPR performance, see Table 3.

Fig. 1) provide a more logical understanding of Sb dynamics and future predictions compared to the linear trends obtained with classical statistics. That is, classical statistics do not allow (i) to obtain reliable Sb_d and Sb_p concentrations for intermediate instants in time between sampled dates (e.g., interpolations), (ii) nor to predict future values in the aquatic system, as the current decreasing trends are assumed to remain constant in the future (which need not be the case), implying that at some point there will be a time where there is no more Sb in the system (i.e., not suitable model for predictions). In fact, it is not the first time that GPR has shown to be an improved tool for time series prediction compared to linear regression (e.g., Camps-Valls et al., 2016). However, out of the literature using GPR in aquatic systems, none of the cited works (Bonakdari et al., 2019; Ding et al., 2023; Gheibi et al., 2022) show the GPR predicted timeseries as we do (e.g., Fig. 1 and Supplementary Fig. S6), hampering further discussions about the general potential of GPR performance in aquatic systems.

4.1.2. More precise seasonal quantification

Classical statistics showed that the seasonal factor (SF, Supplementary Fig. S4 and S5) had recurrent patterns along the year for both Sb_d and Sb_p in all sites (except BP). Overall, Sb_d concentrations were higher than the average concentrations in August and lower in February (Supplementary Fig. S4) whereas for Sb_p it was highest in January and lowest in July (Supplementary Fig. S5). This detailed information geochemically corresponds to the dilution effect of the seasonal variations of Q along the year. However, GPR detects yearly periodicities for all variables except generally for SPM and for Sb_p/Th_p at PSM, T and BP (Table 3). Decoupled periodicities between Q and SPM at all sites except RM, suggest a random component for the SPM and Sb_p/Th_p dynamics from the common Q and Sb_d seasonality (i.e., a detachment between erosion and Q dynamics on a regular basis, potentially related to the influence of dams) particularly for large watersheds: Lot (BP - T) and Garonne (PSM - LR) Rivers vs the smaller Riou Mort River (RM).At RM, common periodicities between Q and SPM show linked hydro-sedimentary dynamics, which influence Sb temporal trends at a local scale, only characterized by the GPR model. The overall independent character of Sb_d and Sb_p/Th_p dynamics could be related to the impact of biological activities, as suggested in Gil-Díaz et al. (2018) for the Lot River, and/or to Sb solid-liquid partitioning (Kd) kinetics (e.g., Sb_p discharge and dissolution is not retraced in the Sb_d dynamics). Thus, despite classical statistics finding a reasonable fit for the Sb_p seasonality, GPR shows little periodic correlation, which is a more logical outcome given the seasonal character of the Kd along the year (*c.f.,* Section 4.2.1). All these outcomes already provide further information and improve the interpretation of the geochemical system compared to the original results from classical statistics (e.g., Gil-Díaz et al., 2018).

4.2. Environmental implications

4.2.1. Estimating solid-liquid partitioning (Kd)

The solid-liquid partitioning coefficient (Kd, L kg^{-1}) indicates the degree of affinity of an element for the particulate phase, thus it is defined as the Sb_p/Sb_d ratio for the case of this work. Field-based Kd values are assumed to reflect the environmental thermodynamic equilibrium of the target element in the system (Sung, 1995). For this reason, Kd values are generally used as reference values for geochemical dispersion models to predict trace element fate in case of accidental releases to the environment or changes in the hydrodynamic regime (Laissaoui et al., 1998; Zheleznyak et al., 2022). In many cases, these values are obtained from onsite, point field campaigns, or from tabulated generalized values in guidelines (TRS 422, International Atomic Energy Agency, 2004). Monitoring programs could also fulfil this requirement, not only based on measured data, but also on interpolated data (i.e., between unmeasured sampling dates or predicted estimations), a point that can be easily provided with GPR models like in this study.

In addition to data interpretation, well characterized GPR models allow further data treatment such as (i) data interpolation, due to correct long-term characterization of the time series (*c.f.* Section 4.1.1), and (ii) calculating site-specific Kd values. For instance, the outcomes from the GPR models correctly represent the seasonal character of the Kd values observed for Sb in the Garonne River (e.g., case of LR, Fig. 2). This is an outcome that classical statistics could provide (e.g., after applying Sen's slope defined in Table S4 and the seasonal factor of Fig. S5), given the definition of the Kd ratio, but with lower precision and interannual variability. The validity of the classical statistics approach is lower than that of the GPR approach, given the accuracy of the modelled results (see r^2 in Fig. S4 – S6).

4.2.2. More detailed process identification

Another advantage of the GPR approach is that developed models of temporal series can be combined. For instance, when adding the modelled mean processes at LR for Q and SPM, we obtain a timeseries with two wave-like trends of different period (Supplementary Fig. S7). When creating a GPR model of this new timeseries based on two periodic

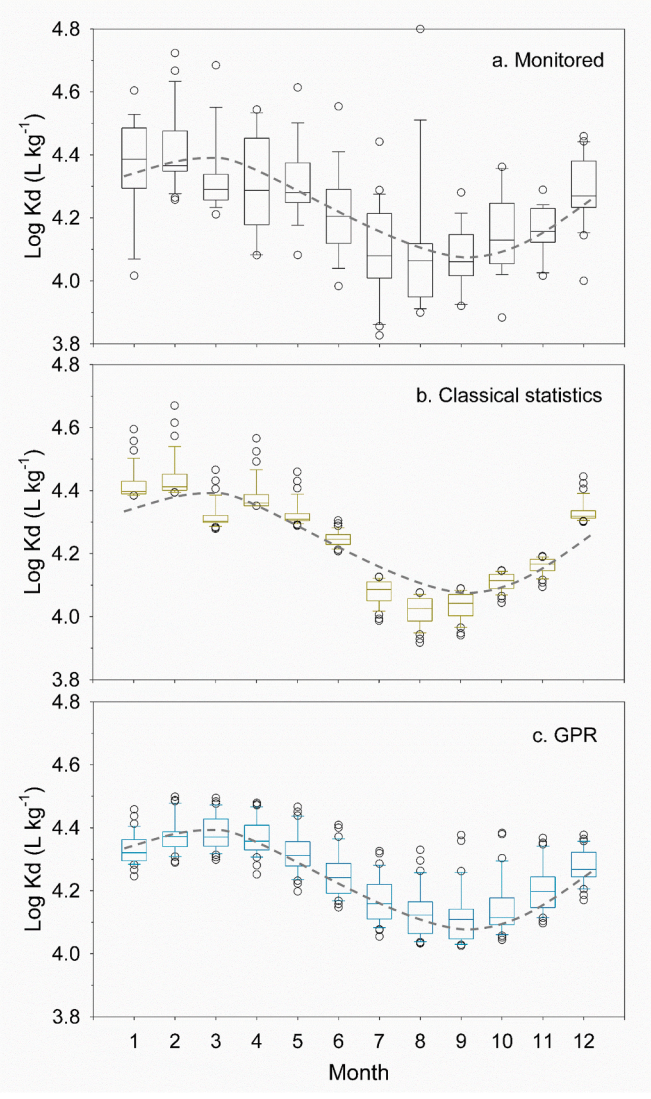


Fig. 2. Comparison between monitored and model derived log Kd values. Results are shown for LR, where the values of solid-liquid partitioning (Kd) were obtained from (a) monitored Sb_d / Sb_p concentrations, (b) extracted Sb_d / Sb_p concentrations from the classical linear decomposition, and (c) extracted Sb_d / Sb_p concentrations from GPR models. Dashed lines are positioned exactly at the same position in each graph and are there to guide the eye.

kernels (i.e., two seasonal components), newly optimized, independent parameters show a long-term period of 4.9 y additionally to the yearly short-term period (Supplementary Fig. S7). One could hypothesize that this long-term periodicity corresponds to a signal from atmospheric-hydrological coupling processes. In fact, when applying similar models to oscillation indices such as the North Atlantic Oscillation (NAO, NAO, 2024) or the Oceanic Niño Index (i.e., ONI, a proxy for classifying El Niño vs La Niña events in the eastern tropical Pacific; ONI, 2024), we obtain long-term periodicities of ~6.2 y, i.e., different to those observed in our dataset (Table 3). Therefore, Q + SPM series in the Garonne watershed might reflect climatic patterns affecting watershed erosion/rainfall-related processes which are not directly linked to the ONI. These results, together with the inconsistent long-term trends for the Sb series at all sites, discourage the potential dominance of atmospheric deposition influencing Sb dynamics in the Garonne watershed, and support the idea of a more hydrological-dominated system, a concept that had not been evidenced until now.

4.3. Applications of GPR models to other geochemical studies in aquatic systems

This work has provided supporting evidence on the robustness and advantages of GPR model performance compared to linear decomposition for the data treatment of geochemical variables in a specific watershed. Nevertheless, the GPR approach can be extended to any temporal series in aquatic systems and its applications are wide, highlighted but not exclusive to the following points:

- *GPR models for system resilience and mass balance characterization*. Well characterized GPR models allow completing in a robust and reliable manner datasets which are measured at low frequency. For example, the GPR model can calculate missing Sb concentrations between the 24-day sampling times, providing more accurate estimations and comprehensible parameters than current approaches that compartmentalize time series and work with averages (Pougnet et al., 2019) or use modern neural networks (e.g., Vu et al., 2022). These estimations with GPR models would allow calculating more accurate daily fluxes of trace elements, providing more precise system mass balance evaluations (net vs gross fluxes). Such calculations would allow to obtain a fast overview on the system resilience (geogenic vs anthropogenic impacts, e.g., Pougnet et al., 2022), improving and advancing at the same time on the knowledge of the geochemical cycles of trace elements, particularly relevant for less studied elements (e.g., Mitra and Sen, 2017). This approach could and should be applied to current monitoring programs worldwide.
- *GPR models for input variables in predictive geochemical models.* Kd values are used as a reference parameter for geochemical dispersion models in post-accidental management studies to predict the fate of trace elements in a given aquatic environment (Laissaoui et al., 1998; Zheleznyak et al., 2022). In many occasions, Kd values are used in an indiscriminate manner, as they either come from a one-time onsite measurement (assuming stationary systems, International Atomic Energy Agency, 2004), from another aquatic system (e.g., Kd values from quality guidelines applied to the Eastern Mediterranean Sea, Tsabaris et al., 2022) or even from another trace element, as suggested for geochemical pairs by certain pre-established guidelines (e.g., TRS 422, International Atomic Energy Agency, 2004). However, Kd values depend on several environmental factors and its magnitude can vary for a given element (e.g., Tomczak et al., 2019). The use of GPR models on local datasets would provide with a simple calculation a better estimation of the element-dependent and local-dependent Kd values, improving the reliability of the geochemical outputs for watershed management. In the presence of complementary SPM datasets, GPR models would additionally help identifying which trace elements and aquatic systems suffer from particle concentration effects (e.g., Zhu et al., 2022), providing more sound knowledge on the use of generalized or local Kd values for decision making.
- *GPR models for atmosphere-hydrosphere coupling dynamics*. Trace element dynamics in aquatic systems may be linked or impacted by atmospheric patterns (Climate Change, atmospheric oscillations, etc.). However, evidence of the influence of atmospheric oscillations and teleconnections on river discharges and/or precipitations are generally shown via correlation coefficients (e.g., Pociask-Karteczka, 2006; Xiao et al., 2015), which is a less powerful approach to GPR. Additionally, the predictive character of the ENSO has been attempted through deep machine learning techniques (e.g., Derot et al., 2024), whereas combinations of GPR models could provide a more interpretable and alternative approach for future studies. Thus, our study (as shown in Section 4.2.2) provides a generalized approach and an example to follow in future works of aquatic systems for simple identification of atmosphere-hydrosphere coupling dynamics.

5. Conclusions

In this work, GPR has proven to be a powerful tool for both time series analysis and understanding of geochemical data. That is, GPR can help evaluate the effect and limitations of environmental sampling frequencies on the predictive power of a model from a given time series. Our systematic analysis on water discharge time series showed that dataset frequency rather than dataset length plays a greater role for accurately characterizing the aquatic system via GPR models. This means that, GPR is a robust model that can identify trends and seasonal patterns from time series of variable lengths, allowing to have a deeper insight of aquatic geochemical patterns/correlations which are not evident to other non-parametric approaches. Once temporal series of geochemical data are characterized, they can be used to derive other parameters such as trace element concentrations and the solid/liquid partitioning of trace elements, often required for system mass balance calculations and used as inputs for post-accidental management strategies. Overall, the use of GPR models can be extended to any temporal series and its applications are wide, providing a more accurate approach than simple, linear compartmentalization of the time series. Future studies will explore the potential of combining GPR models to further understand inter-site dependence and/or intercorrelations between the dynamics of different geochemical parameters in aquatic systems.

CRediT authorship contribution statement

Teba Gil-Díaz: Writing – review & editing, Writing – original draft, Visualization, Supervision, Resources, Project administration, Investigation, Funding acquisition, Data curation. **Michael Trumm:** Writing – review & editing, Writing – original draft, Visualization, Software, Resources, Methodology, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare no competing interests.

Acknowledgements

This work was supported by the Federal Ministry of Education and Research (BMBF) and the Baden-Württemberg Ministry of Science as part of the Excellence Strategy of the German Federal and State Governments. The authors also acknowledge support by the High Performance and Cloud Computing Group at the Zentrum für Datenverarbeitung of the University of Tübingen, the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant n° INST 37/935-1 FUGG.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2024.102877.

Data availability

Raw data is attached as Supplementary in an excel file.

References

Ballabio, C., Lugato, E., Fernández-Ugalde, O., Orgiazzi, A., Jones, A., Borrelli, P., Montanarella, L., Panagos, P., 2019. Mapping LUCAS topsoil chemical properties at European scale using Gaussian process regression. Geoderma 35, 113912. https://doi.org/10.1016/j.geoderma.2019.113912.

Bonakdari, H., Ebtehaj, I., Samui, P., Gharabaghi, B., 2019. Lake water-level fluctuations forecasting using minimax probability machine regression, relevance vector machine, Gaussian process regression, and extreme learning machine. Water Resour. Manag. 33, 3965–3984. https://doi.org/10.1007/s11269-019-02346-0.

Camastra, F., Casolaro, A., Iannuzzo, G., 2023. Manifold learning by a deep Gaussian process autoencoder. Neural Comput. & Applic. 35 (21), 15573–15582. https://doi.org/10.1007/s00521-023-08536-7.

Camps-Valls, G., Verrelst, J., Munoz-Mari, J., Laparra, V., Mateo-Jimenez, F., Gomez-Dans, J., 2016. A survey on Gaussian processes for earth-observation data analysis: a comprehensive investigation. IEEE Geosci. Remote Sens. Magaz. 4 (2), 58–78. https://doi.org/10.1109/MGRS.2015.2510084.

Camps-Valls, G., Sejdinovic, D., Runge, J., Reichstein, M., 2019. A perspective on Gaussian processes for earth observation. Natl. Sci. Rev. 6, 616–618. https://doi.org/10.1093/nsr/nwz028.

Coynel, A., Schäfer, J., Hurtrez, J.-E., Dumas, J., Etcheber, H., Blanc, G., 2004. Sampling frequency and accuracy of SPM flux estimates in two contrasted drainage basins. Sci. Total Environ. 330, 233–247. https://doi.org/10.1016/j.scitotenv.2004.04.003.

Derot, J., Sugiura, N., Kim, S., Kouketsu, S., 2024. Improved climate time series forecasts by machine learning and statistical models coupled with signature method: a case study with El Niño. Ecol Inform. 79, 102437. https://doi.org/10.1016/j.ecoinf.2023.102437.

Di, Z., Chang, M., Guo, P., 2019. Water quality evaluation of the Yangtze River in China using machine learning techniques and data monitoring on different time scales. Water 11 (2), 339–362. https://doi.org/10.3390/w11020339.

Ding, C., Sun, S., Zhao, J., 2023. MST-GAT: a multimodal spatial–temporal graph attention network for time series anomaly detection. Inform Fusion 89, 527–536. https://doi.org/10.1016/j.inffus.2022.08.011.

Domingues, R., Michiardi, P., Zouaoui, J., Filippone, M., 2018. Deep Gaussian process autoencoders for novelty detection. Mach. Learn. 107, 1363–1383. https://doi.org/10.1007/s10994-018-5723-3.

Elbeltagi, A., Azad, N., Arshad, A., Mohammed, S., Mokhtar, A., Pande, C., Etedali, H.R., Bhat, S.A., Islam, ArMdT, Deng, J., 2021. Applications of Gaussian process regression for predicting blue water footprint: case study in ad Daqahliyah, Egypt. Agr. Water Manage. 255, 107052. https://doi.org/10.1016/j.agwat.2021.107052.

Gheibi, M., Eftekhari, M., Akrami, M., Emrani, N., Hajiaghaei-Keshteli, M., Fathollahi-Fard, A.M., Yazdani, M., 2022. A sustainable decision support system for drinking water systems: resiliency improvement against cyanide contamination. Infrastructures 7, 88–106. https://doi.org/10.3390/infrastructures7070088.

Gil-Díaz, T., Schäfer, J., Coynel, A., Bossy, C., Dutruch, L., Blanc, G., 2018. Antimony in the Lot–Garonne River system: a 14-year record of solid–liquid partitioning and fluxes. Environ. Chem. 15, 121–136. https://doi.org/10.1071/EN17188.

Hanson, P.C., Stillman, A.B., Jia, X., Karpatne, A., Dugan, H.A., Carey, C.C., Stachelek, J., Ward, N.K., Zhang, Y., Read, J.S., Kumar, V., 2020. Predicting lake surface water phosphorus dynamics using process-guided machine learning. Ecol. Model. 430, 109136. https://doi.org/10.1016/j.ecolmodel.2020.109136.

Hoolohan, V., Tomlin, A.S., Cockerill, T., 2018. Improved near surface wind speed predictions using Gaussian process regression combined with numerical weather predictions and observed meteorological data. Renew. Energy 126, 1043–1054. https://doi.org/10.1016/j.renene.2018.04.019.

Houston, J.F.T., 1983. Ground-water systems simulation by time-series techniques. Groundwater 21, 301–310. https://doi.org/10.1111/j.1745-6584.1983.tb00729.x.

Hu, J., Wang, P., Li, D., Liu, S., 2024. A long-term multivariate time series prediction model for dissolved oxygen. Ecol Inform. 82, 102695. https://doi.org/10.1016/j.ecoinf.2024.102695.

International Atomic Energy Agency, 2004. Sediment Distribution Coefficients and Concentration Factors for Biota in the Marine Environment. Technical Reports Series, vol. 422. IAEA, Vienna.

Khalil, B., Ouarda, T.B.M.J., 2009. Statistical approaches used to assess and redesign surface water-quality-monitoring networks. J. Environ. Monit. 11, 1915–1929. https://doi.org/10.1039/B909521G.

Kovač-Andrić, E., Brana, J., Gvozdić, V., 2009. Impact of meteorological factors on ozone concentrations modelled by time series analysis and multivariate statistical methods. Ecol Inform. 4 (2), 117–122. https://doi.org/10.1016/j.ecoinf.2009.01.002.

Laissaoui, A., Abril, J.M., Periáñez, R., León, M.G., Montaño, E.G., 1998. Kinetic transfer coefficients for radionuclides in estuarine waters: Reference values from 133 Ba and effects of salinity and suspended load concentration. J. Radioanal Nucl. Ch. 237, 55–62. https://doi.org/10.1007/BF02386662.

Lin, C., Li, T., Chen, S., Liu, X., Lin, C., Liang, S., 2019. Gaussian process regression-based forecasting model of dam deformation. Neural Comput. & Applic. 31, 8503–8518. https://doi.org/10.1007/s00521-019-04375-7.

Liu, H., Ong, Y.S., Shen, X., Cai, J., 2020. When Gaussian process meets big data: a review of scalable GPs. IEEE Trans. Neural Netw. Learn. Syst. 31 (11), 4405–4423. https://doi.org/10.1109/TNNLS.2019.2957109.

Matthews, A.G.G., van der Wilk, M., Nickson, T., Fujii, K., Boukouvalas, A., León-Villagrá, P., Ghahramani, Z., Hensman, J., 2017. GPflow: a Gaussian process library using TensorFlow. J. Mach. Learn. Res. 18, 1–6. https://doi.org/10.48550/arXiv.1610.08733.

Mitra, A., Sen, I.S., 2017. Anthrobiogeochemical platinum, palladium and rhodium cycles of earth: emerging environmental contamination. Geochim. Cosmochim. Ac. 216, 417–432. https://doi.org/10.1016/j.gca.2017.08.025.

NAO, 2024. North Atlantic Oscillation Dataset from NOAA. http://www.cpc.ncep.noaa.gov/products/precip/CWlink/pna/nao.shtml. Last accessed: 13th February 2024.

Nielsen, A., 2019. Practical Time Series Analysis: Prediction with Statistics and Machine Learning. O'Reilly Media.

Nortier, B., Williamson, D., Mancini, M., Bateman, I., Day, B., Binner, A., 2024. Deep gaussian processes and inversion for decision support in model-based climate change mitigation and adaptation problems. In: International Conference on Learning Representation. Accessible in. https://s3.us-east-1.amazonaws.com/climate-change-ai/papers/iclr2024/59/paper.pdf [Last accessed on 18/10/2024].

ONI, 2024. Oceanic Niño Index Dataset from NOAA. https://ggweather.com/enso/oni.htm. Last accessed: 13th February 2024.

Passalis, N., Tefas, A., Kanniainen, J., Gabbouj, M., Iosifidis, A., 2020. Deep adaptive input normalization for time series forecasting. IEEE T. Neur. Net. Lear. 31, 3760–3765. https://doi.org/10.1109/TNNLS.2019.2944933.

Pelé, K., Nicoulaud-Gouin, V., Lepage, H., 2024. A neural network encoder-decoder for time series prediction: application on ^{137}Cs particulate concentrations in nuclearized rivers. Ecol Inform. 80, 102463. https://doi.org/10.1016/j.ecoinf.2024.102463.

Pociask-Karteczka, J., 2006. River hydrology and the North Atlantic oscillation: a general review. Ambio 312–314. DOI: jstor.org/stable/4315743.

Pougnet, F., Blanc, G., Mulamba-Guilhemat, E., Coynel, A., Gil-Diaz, T., Bossy, C., Strady, E., Schäfer, J., 2019. New computation for a better estimation of the annual dissolved metal net fluxes. The case of the cadmium in the Gironde estuary. Hydroécol. Appl. https://doi.org/10.1051/hydro/2019002.

Pougnet, F., Gil-Díaz, T., Blanc, G., Coynel, A., Bossy, C., Schäfer, J., 2022. Historical mass balance of cadmium decontamination trends in a major European continent-ocean transition system: case study of the Gironde Estuary. Mar. Environ. Res. 176, 105594. https://doi.org/10.1016/j.marenvres.2022.105594.

Ranjbar, M.H., Etemad-Shahidi, A., Kamranzad, B., 2020. Modeling the combined impact of climate change and sea-level rise on general circulation and residence time in a semi-enclosed sea. Sci. Total Environ. 740, 140073. https://doi.org/10.1016/j.scitotenv.2020.140073.

Rasmussen, C.E., Williams, C.K., 2006. Gaussian Processes for Machine Learning, vol. 1. MIT Press, Cambridge. https://doi.org/10.7551/mitpress/3206.001.0001.

Robson, B.J., 2014. When do aquatic systems models provide useful predictions, what is changing, and what is next? Environ. Model. Softw. 61, 287–296. https://doi.org/10.1016/j.envsoft.2014.01.009.

Sajib, A.M., Diganta, M.T.M., Moniruzzaman, M., Rahman, A., Dabrowski, T., Uddin, M. G., Olbert, A.I., 2024. Assessing water quality of an ecologically critical urban canal incorporating machine learning approaches. Ecol Inform. 80, 102514. https://doi.org/10.1016/j.ecoinf.2024.102514.

Sung, W., 1995. Some observations on surface partitioning of Cd, Cu and Zn in estuaries. Environ. Sci. Technol. 29, 1303–1312. DOI: 0013-936X/95/0929-1303.

Tomczak, W., Boyer, P., Krimissa, M., Radakovitch, O., 2019. Kd distributions in freshwater systems as a function of material type, mass-volume ratio, dissolved organic carbon and pH. Appl. Geochem. 105, 68–77. https://doi.org/10.1016/j.apgeochem.2019.04.003.

Tsabaris, C., Eleftheriou, G., Tsiaras, K., Triantafyllou, G., 2022. Distribution of dissolved ^{137}Cs, ^{131}I and ^{238}Pu at eastern Mediterranean Sea in case of hypothetical accident at the Akkuyu nuclear power plant. J. Environ. Radioact. 251, 106964. https://doi.org/10.1016/j.jenvrad.2022.106964.

Vu, Q., 2023. Deep Statistical Models with Application to Environmental Data, Doctor of Philosophy Thesis, School of Mathematics and Applied Statistics. University of Wollongong. https://ro.uow.edu.au/theses1/1632.

Vu, H.L., Ng, K.T.W., Richter, A., Li, J., Hosseinipooya, S.A., 2022. Impacts of nested forward validation techniques on machine learning and regression waste disposal time series models. Ecol Inform. 72, 101897. https://doi.org/10.1016/j.ecoinf.2022.101897.

Wunsch, A., Liesch, T., Broda, S., 2022. Deep learning shows declining groundwater levels in Germany until 2100 due to climate change. Nat. Commun. 1-13, 1221. https://doi.org/10.1038/s41467-022-28770-2.

Xiao, M., Zhang, Q., Singh, V.P., 2015. Influences of ENSO, NAO, IOD and PDO on seasonal precipitation regimes in the Yangtze River Basin, China. Int. J. Climatol. 35 (12), 3556–3567. https://doi.org/10.1002/joc.4228.

Zhang, Z., Wang, C., Peng, X., Qin, H., Lv, H., Fu, J., 2021. Solar radiation intensity probabilistic forecasting based on K-means time series clustering and Gaussian process regression. IEEE Access 9, 89079–89092. https://doi.org/10.1109/ACCESS.2021.3077475.

Zheleznyak, M., Kivva, S., Pylypenko, O., Sorokin, M., 2022. Chapter 11. Modeling of behaviour of Fukushima-derived radionuclides in freshwater systems. In: Nanba, K.,

Konoplev, A., Wada, T. (Eds.), Behavior of Radionuclides in the Environment III: Fukushima. Springer Nature.
Zhu, S., Piotrowski, A.P., 2020. River/stream water temperature forecasting using artificial intelligence models: a systematic review. Acta Geophys. 68, 1433–1442. https://doi.org/10.1007/s11600-020-00480-7.
Zhu, J., Xu, C., Chen, C., Zhang, A., Shao, J., Zhang, Q., 2022. Solution to the particle concentration effect on determining Kd value of radionuclides. J. Environ. Radioact. 255, 107028. https://doi.org/10.1016/j.jenvrad.2022.107028.