International Journal of Environmental Research and Public Health

MDPI

Article

Reconstructing missing and anomalous data collected from high-frequency in-situ sensors in fresh waters

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

[1] CNRS/Univ Pau & Pays Adour, Laboratoire de Mathématiques et de leurs Applications de Pau Fédération MIRA, UMR 5142, Anglet 64600, France
[2] Department of Mathematics and Statistics, Macquarie University, Sydney 2109, New South Wales, Australia
[3] National Ecological Observatory Network, Battelle Boulder 80301, Colorado, United States of America
[4] Institute of Arctic Biology, and Department of Biology and Wildlife, University of Alaska Fairbanks, Fairbanks 99775, Alaska, United States of America
[5] School of Mathematical Sciences, Queensland University of Technology, Brisbane 4000, Queensland, Australia
[6] ARC Centre of Excellence for Mathematics and Statistical Frontiers, Australia
[7] Peterson Consulting, Brisbane 4000, Queensland, Australia
[8] Monash University, Clayton 3800, Victoria, Australia
[9] Biosciences and Food Technology Discipline, School of Science, RMIT University, Bundoora 3083, Victoria, Australia
* Correspondence claire.kermorvant@univ-pau.fr

Abstract: In-situ sensors that collect high-frequency data are used increasingly to monitor aquatic environments. These sensors are prone to technical errors, resulting in unrecorded observations and/or anomalous values that are subsequently removed and create gaps in time series data. We present a framework based on generalized additive and auto-regressive models to recover these missing data. To mimic sporadically missing (i) single observations and (ii) periods of contiguous observations, we randomly removed (i) point data and (ii) day and week-long sequences of data from a two-year time series of nitrate-concentration data collected from Arikaree River, USA, where synoptically collected water temperature, turbidity, conductance, elevation and dissolved oxygen data were available. In 72% of cases with missing point data, predicted values were within the sensor-precision interval of the original value, although predictive ability declined when sequences of missing data occurred. Precision also depended on the availability of other water-quality covariates. When covariates were available, even a sudden, event-based peak in nitrate concentration was reconstructed well. By providing a promising method for accurate prediction of missing data, the utility and confidence in summary statistics and statistical trends will increase, thereby assisting the effective monitoring and management of fresh waters and other at-risk ecosystems.

Keywords: Anomaly correction; Generalised additive model (GAM); Missing data reconstruction; Remote sensing; Water quality

Citation: Kermovant, C., et al. Reconstructing missing and anomalous data collected from high-frequency in-situ sensors in fresh waters. *Int. J. Environ. Res. Public Health* **2021**, *18*, x. https://doi.org/10.3390/xxxxx

Academic Editor: Firstname Lastname

Received: date
Accepted: date
Published: date

1. Introduction

Water-quality sampling and analysis commonly relies on manual approaches such as grab sampling and laboratory analyses, often conducted at monthly or longer intervals for variables such as sediment and nutrient concentration [1]. As such, the ability to capture water-quality events or determine patterns and trends at fine spatial and temporal resolution are often limited [2]. Advances in the development of in-situ, high-frequency environmental sensors are leading to their expanded use in environmental monitoring, including for fresh waters [3]. As the cost-effectiveness and telecommunications capability

of these sensors increases, their ability to provide high frequency data in near-real time likewise increase, allowing managers and decision makers to act in a timelier and more spatially specific fashion. The large datasets generated by high frequency in-situ sensors also present new opportunities for scientists when analysing, modelling and reporting water-quality data [4, 5]. Consequently, high-frequency datasets collected from in-situ sensors can provide a more thorough understanding water-quality dynamics at multiple time scales and help to improve data quality assurance and quality control [6].

In-situ sensors, despite their benefits, are prone to technical errors due to biofouling, power failures and other issues. These errors can lead to technical anomalies in water-quality data and potentially confound the assessment or identification of true changes in water chemistry [7]. Given that the high frequency and large size of these datasets precludes the use of manual anomaly-detection methods (one part of the entire data quality assurance and quality control process), various automated approaches have been proposed. For example, Shi et. al (2018) [8] integrated a wavelet artificial neural network with surrogate measurements for rapid warning of water-quality anomalies, Liu et. al. (2020) [9] integrated a Bayesian autoregressive model with an Isolation Forest algorithm for combined prediction and detection, while Rodriguez-Perez et. al. (2020) [10] developed a semi-supervised, Bayesian artificial neural network approach. To assist both developers and end-users, Leigh et. al. (2019) [7] developed a ten-step anomaly detection framework to systematically implement and compare suites of anomaly detection methods based on end-user needs.

Regardless of the method used to detect water-quality anomalies from in-situ sensors, observations that get labelled as anomalous are often removed subsequently from the time series, rendering them missing. Furthermore, given the variety of types of technical anomalies, such as sudden spikes, unrealistic values, drift or periods of anomalously high or low variability [7], the resultant time series may contain missing point observations and/or sequences of contiguously missing observations after the data are passed through an anomaly detection algorithm. Failure to replace anomalies with corrected data may occur because methods to confidently reconstruct (accurately predict) the true values of the missing water-quality observations are not available. Missing data then create data quality issues [11] and can lead to biased estimates of parameters, increased standard errors, decreased statistical power, and lost information [12] which may hinder the calculation of summary statistics [13] and affect statistical trend detection [14]. Slater et al., 2017 [15], for example, demonstrate via simulation that the loss in trend detection tends to increase with increasing size of missing data 'gaps' and decreasing length of time series.

Many of the commonly used methods used to reconstruct missing water-quality data were developed before the proliferation of high-frequency sensors, such as infilling based on surrounding data [16, 17], regression analysis [18, 19], state-space models with Estimation-Maximization algorithm [20] or artificial neural networks [21, 22, 23], and therefore were targeted at data with lower-frequency time steps, such as daily data. More recently, but also based on daily data, various infilling techniques such as regression, scaling and equi-percentile approaches [24], along with dynamic regression models [25] have been used to reconstruct missing streamflow data. Methods developed in other domains, such as computer science, have reconstructed missing sensor data based on temporal or spatial correlation, interpolation and sparse theory [26]. In sensor networks, linear and non-linear regression methods have been developed that use the non-missing data adjacent data to the missing datum [27], along with algorithms based on combining K-means algorithms and neural networks with particle swarm optimization [28]. Similar methods have been developed in other domains, such as those used for power systems within computer science [29], while in the engineering domain, bidirectional recurrent neural networks have been developed to reconstruct sensor data used to monitor bridge construction [30]. As can be seen, these various methods of data reconstruction have been developed fit for purpose and as solutions for domain-specific problems. Hence, we

aimed to develop a suitable method to reconstruct high-frequency nutrient data collected from in-situ sensors in rivers, a problem to our knowledge that is yet to be addressed.

In the environmental domain, and specifically river management, nutrient monitoring and specifically that of nitrate concentration is particularly important. In its bio-available form, nitrate is assimilated for growth and metabolism by riverine biota (e.g., algae, macrophytes and some bacteria) that form the basal components of aquatic food webs [31]. However, an excess of nitrate can lead to problems like eutrophication, leading to decrease in light infiltration and dissolved oxygen concentration [32, 33], which in turn can negatively affect the health of aquatic biota such as fishes and invertebrates [34, 35, 36], as well as increasing costs for water treatment and complicating management of river ecosystems spanning catchment headwaters to receiving waters downstream, including oceans [37]. Furthermore, nitrate concentration can vary substantially in space and time in river ecosystems due to instream processes and external inputs [38, 39]. This high spatial and temporal variation has increased the interest in and use of high-frequency, in-situ nitrate sensors in river monitoring programs and thus the need to develop appropriate methods to reconstruct missing nitrate-concentration data from the resulting time series.

While it is important to develop a sound method to confidently reconstruct missing nitrate data for use in environmental management, the use of nitrate data can also serve as a case study to demonstrate the potential for the method to be applied more broadly. As such, the objective of this study was to develop and test a data reconstruction method using both a real time series of high-frequency nitrate concentrations and a simulation study.

2. Materials and Methods

2.1. Reconstruction method

Let Y be the response (i.e., dependent) variable of interest and Y_t the value taken by Y at time t. For covariates $X \in m$ (i.e., m explanatory variables or predictors) we denote X_{kt}, the kth covariates observed at time t. We then identified two possible cases: (i) all X_{kt} are available at the same time step as the variable of interest Y_t, and (ii) at least one covariate is not available at time t. For the first case, when all X_{kt} are available, we used a generalised additive model to predict Y at each time t that is missing, following the equation:

$$Y_t = \beta_0 + \sum_{k=1}^{m} s_k(X_{kt}) + \varepsilon_t \quad (1)$$

where X_{kt} are covariates measured at the tth sample. Here, β_0 is an intercept and ε_t is an error term, following the usual i.i.d assumptions that we make about regression errors, $\varepsilon_t \sim N(0, \sigma 2)$. The associated smooth function $s_k(\cdot)$ of each water-quality variable X_k was defined using thin plate spline regression [40]. A forwards and backwards stepwise variable-selection procedure was implemented and the 'best' GAM model (in terms of variables selected and penalisation of smooth splines) was identified based on the Akaike Information Criterion (AIC) [41].

For the case when at least one covariate is not available at time t, such that Y_t could not been predicted with the GAM model, we used an autoregressive integrated moving average model (ARIMA, Figure 1) [42]. For each missing Y_t that could not been predicted with a GAM, we used the 500 previous Y observations (Y_{t-500}, Y_{t-499}, ..., Y_{t-2}, Y_{t-1}) and selected the best ARIMA model by AIC comparisons. Prediction intervals (95%) associated with the reconstructed values were then calculated according to the model used.

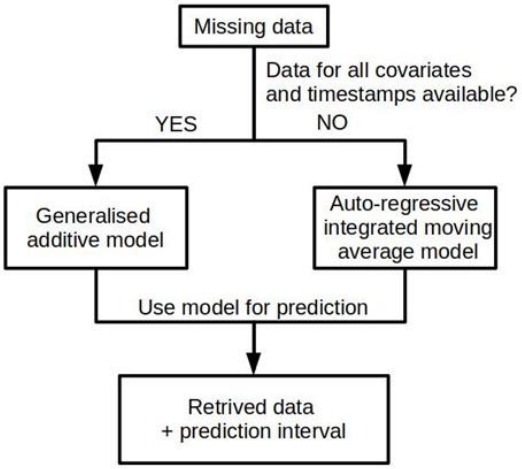

Figure 1. Reconstruction method. Flow chart of the method to predict the values of missing observations in high-frequency sensor data.

2.2. *Arikaree River case study: applying the reconstruction method*

We applied the missing-data reconstruction method to time series of water-quality data collected from Arikaree River, a small wadeable stream in the semi-arid Eastern Colorado plains of the United States of America. The Arikaree River site has a catchment of 2,632 km² comprising mainly grasslands and irrigated agricultural and is part of the National Ecological Observatory Network (NEON). NEON collects and provides open data from aquatic and terrestrial sites across the United States of America (USA), including data from high-frequency, in-situ sensors. NEON conducts standardised configuration, calibration and preventive maintenance procedures on all their sensors [43, 44] and follows in-situ measurement and sample analysis protocols as outlined in [45]. As such, the Arikaree River site provided us with suitable time-series of water-quality data for the purposes of this study.

Several water-quality variables are available from each NEON aquatic site (Table A1). Nitrate concentration [46] is measured in µmol/L using a 10-mm path length SUNA V2 UV light spectrum sensor. The SUNA V2 collects data reported as a mean value from 20 measurements made during a sampling burst every 15 minutes. The published nitrate resolution is 0.1 µmol/L and the manufacturer's stated sensor accuracy is approximately 2 µmol/L or 10% of the reading above 20 µmol/L. We therefore report units of measurement for nitrate in µmol/L (1 µmol nitrate/ L = 0.062 mg nitrate/L). Other co-located sensors report specific conductance (µS/cm), dissolved oxygen (mg/L), water temperature (◦C) and turbidity (Formazin Nephelometric Units, FNU) data as one-minute instantaneous measurements [47, 48]. Water elevation data (i.e., water level as meters above sea level) are published as five-minute averaged measurements from data sampled at 1-minute intervals [49].

For this study, we used a two-year period of nitrate concentration data from one of two sites (the downstream site) on the Arikaree River, from October 2018 to October 2020 (n = 73,056 nitrate concentration observations), in which there were already missing point data and missing periods of data (14,283 missing observations of nitrate concentration = 20% of the nitrate data in total) (Figure 2). These data had all been removed from the time series as part of the NEON data quality assurance and quality control process. For example, there was a technical issue with the nitrate sensor during winter 2019 so no nitrate concentration measurements were available for the first three months of 2019. Missing data were also present in the time series of the covariates as a result of quality control and assurance processing: 29% of the temperature time-series, 35% of the specific

conductance time-series, 14% of the dissolved oxygen time-series, 47% of the turbidity time-series and 11% of the elevation time-series.

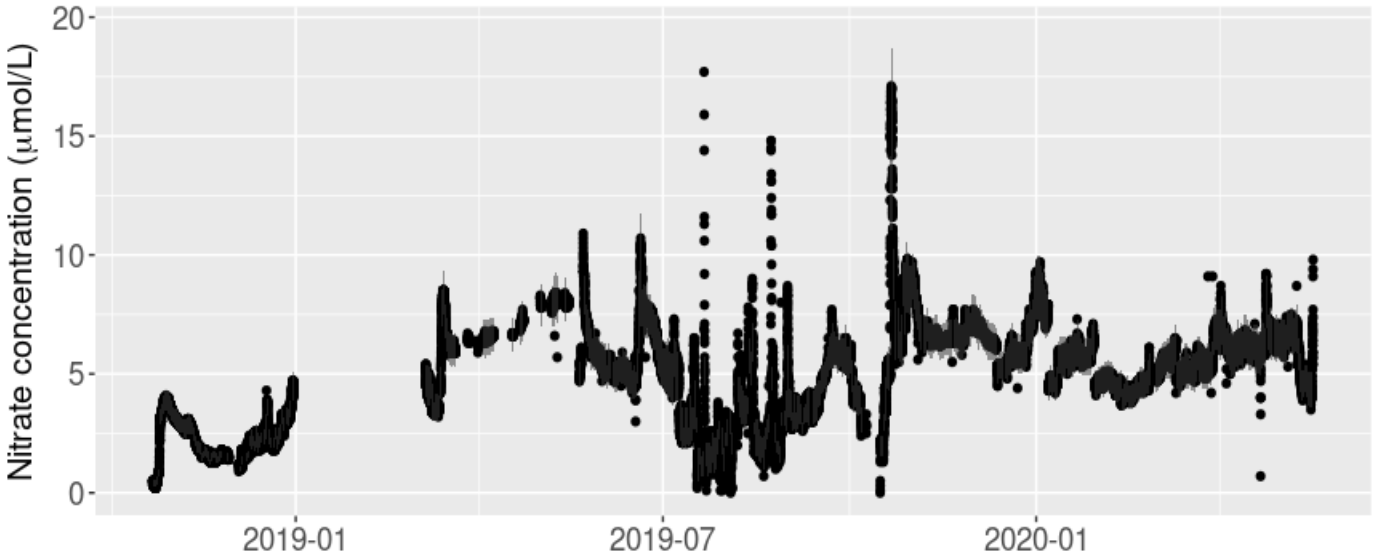

Figure 2. Arikaree River nitrate data. Black points represent the original nitrate observations, gray shading represents the precision interval of the sensor.

We considered nitrate concentration as our Y and the other water-quality variables (specific conductance, dissolved oxygen, water temperature, turbidity and water elevation) as the covariates X that could be related to Y [50]. Visual examination of the distributions of the response and covariates indicated that turbidity had a strongly right-skewed distribution and was therefore log-transformed (i.e. log (turbidity + 1)) prior to analysis [51]. We also included two additional covariates to account for temporal autocorrelation in the time series, as determined by the AIC. The first additional covariate was nitrate concentration at one time-step before time t (i.e. Y_{t-1}) and the second was nitrate concentration at two time-steps before time t (i.e. Y_{t-2}).

2.3. Simulation study: performance evaluation

To evaluate the performance of our reconstruction method, we then repeatedly and randomly removed different combinations of data (both point observations and sequences of contiguous observations) from the two-year time series of nitrate concentration from Arikaree River. For the missing point data, we randomly removed 20%, 30%, and 40% of the observations from the nitrate concentration time series and repeated this process 100 times each (Simulations 1, 2 and 3). For the missing sequences of data, we randomly removed ten individual days (10 x 24 h worth) of observations, repeating the process 100 times (Simulation 4), as well as ten individual weeks of observations, again repeating the process 100 times (Simulation 5).

For each simulation, we then calculated the root-mean-square error (RMSE) and the proportion of reconstructed data within the precision interval of the nitrate sensor (PWPI), i.e., ± 10% for readings > 20 µmol/L and ± 2 µmol/L for readings < 20 µmol/L.

2.4. Implementation

Simulation and imputation were performed with the base packages within the R statistical software [52]. Modelling was undertaken using the car [53], gam [54], mgcv [39] and forecast [55] packages. The R script used to implement the analyses is provided in the GitHub repository available online at https://github.com/Claire-K/nitrate_time_serie_reconstruction and the Arikaree data are available from NEON [46-49] (see Table A1 for data product numbers).

3. Results

3.1. Arikaree River case study

Model performance varied according to the characteristics of the missing data. For demonstration purposes, we focus here on four different cases: (a) a 12-day sequence in which data were sporadically missing, (b) a one-day peak-flow event containing sporadically missing data, (c) a full day of missing data, and (d) a three-month sequence of missing data (Figure 3). Our method performed well at predicting values of nitrate concentration where point observations were sporadically missing from the time series. In other words, the predicted values followed the pattern of the surrounding data closely and prediction intervals were narrow compared to the sensor precision interval (e.g. Figure 3a,b).

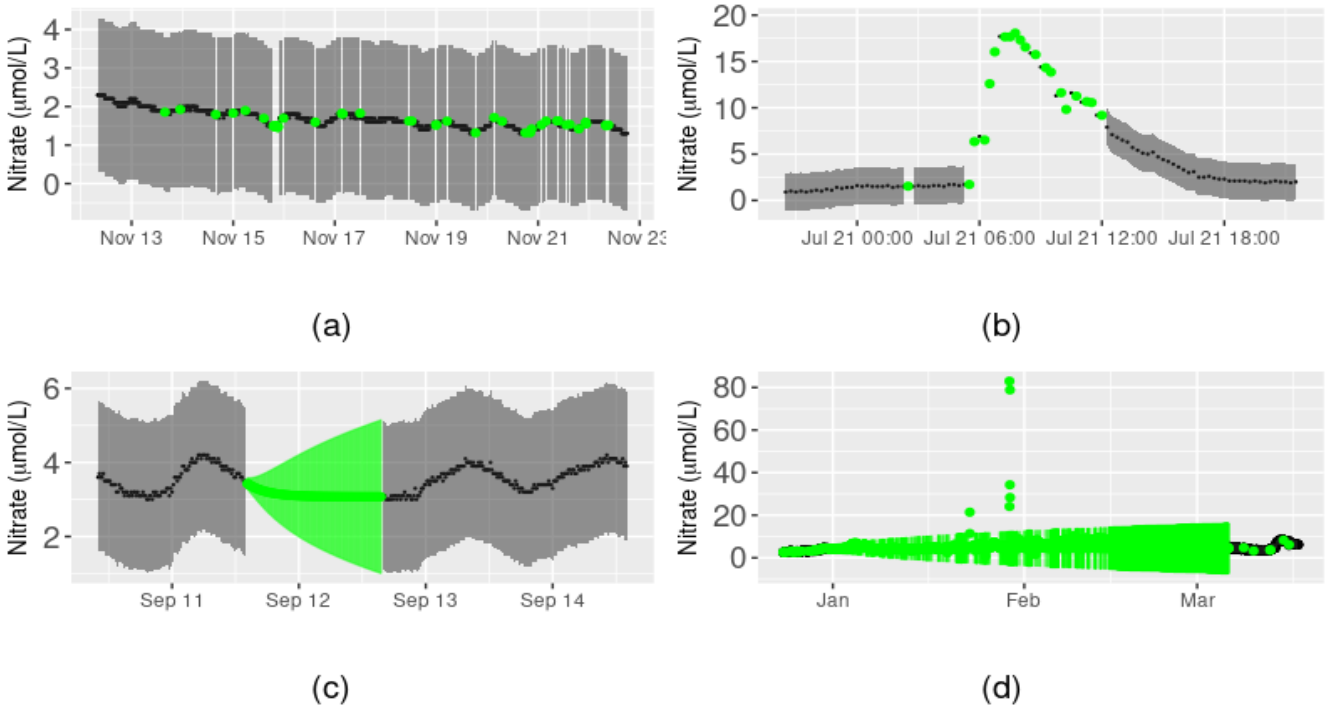

Figure 3. Reconstruction of Arikaree nitrate data. Green points represent the nitrate concentration values predicted by the reconstruction method, along with intervals of prediction, for (a) a 12-day sequence of sporadically missing data, (b) a one-day peak-flow event containing sporadically missing data, (c) a full day of contiguously missing data, and (d) a three-month sequence of contiguously missing data. Black points represent the original nitrate observations, gray shading represents the precision interval of the sensor. Prediction intervals may not be visible when they are narrow relative to the precision interval of the sensor.

However, the method performed less well when periods of contiguously missing observations were reconstructed. For the single day of missing data, the daily pattern in nitrate concentration present in the surrounding data was not reconstructed, and the prediction interval of the reconstructed nitrate values increased with the number of missing observations (Figure 3c). This was also the case for the reconstruction of the three-month period of missing data (Figure 3d). However, some extremely high nitrate concentrations (~ 80 µmol/L) were predicted to occur during this period, based on the values of the covariates at the time, which had not been detected as anomalous by the data quality assurance and control process [56]. This demonstrated that the quality of the reconstructed data can depend heavily on the covariates, when available, and therefore reliable performance of any anomaly detection method implemented prior to reconstruction is crucial.

We also found that reconstructed values of nitrate had much larger prediction intervals when ARIMA rather than GAM was used due to the presence of missing data in

the covariate(s), which simply by chance would more often occur during contiguous sequences of missing nitrate data than during periods of similar length with sporadically missing nitrate data. Overall, the prediction interval for the 14, 283 missing values in the nitrate time series ranged from 0.01 to 56.03 µmol/L, with a median of 1.34 µmol/L.

3.2. *Simulation study: performance evaluation*

3.2.1. Simulations 1, 2 and 3: missing point data

In terms of the reconstruction performance of our method, the RMSE from simulations 1, 2 and 3 (20%, 30% and 40% of randomly missing point data in the nitrate time series, respectively) were all similar and rarely > 0.2 µmol/L, even with 40% of the data having been removed (Figure 4a). Furthermore, the method predicted more than 95% of the missing nitrate values with an RMSE of 0.2 µmol/L. Nevertheless, as the proportion of missing data increased, so did the maximum RMSE. Overall, 72% of the reconstructed nitrate values were within the precision interval of the sensors (Figure 4b).

The performance of our method in reconstructing missing point data can also be demonstrated by looking more closely at different periods of the simulated Arikaree River time series, including typical baseflow and storm-event behaviors of nitrate concentration. In all cases, the predicted data followed the pattern of nitrate concentration closely, including a peak event that occurred over a period of less than 24 h (Figure 5).

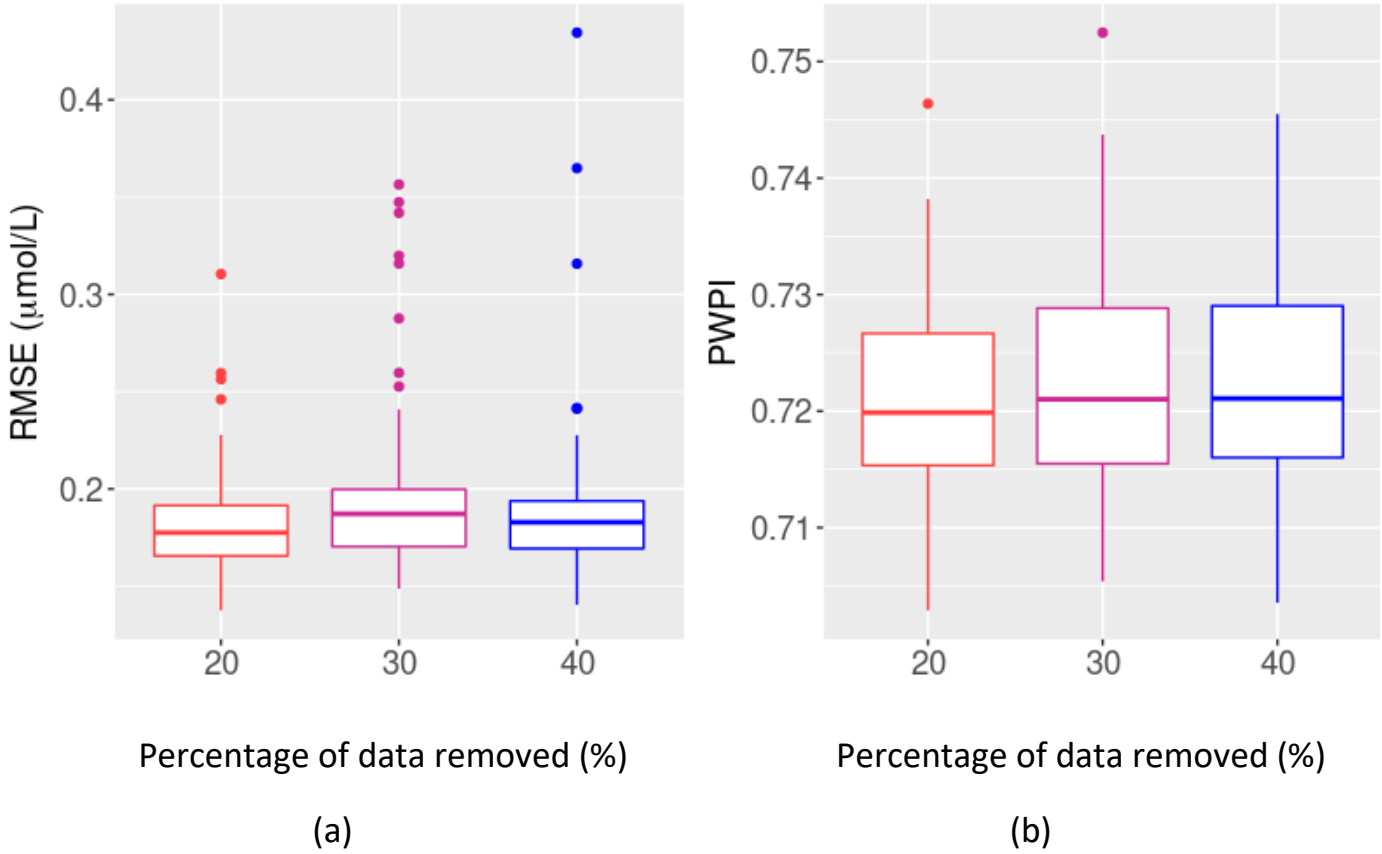

Figure 4. Performance evaluation: reconstructing missing point data. Boxplots of (**a**) root-mean-squared-error (RMSE) and (**b**) the proportion of reconstructed data within the precision interval (PWPI) for different amounts of randomly removed point observations, simulated 100 times.

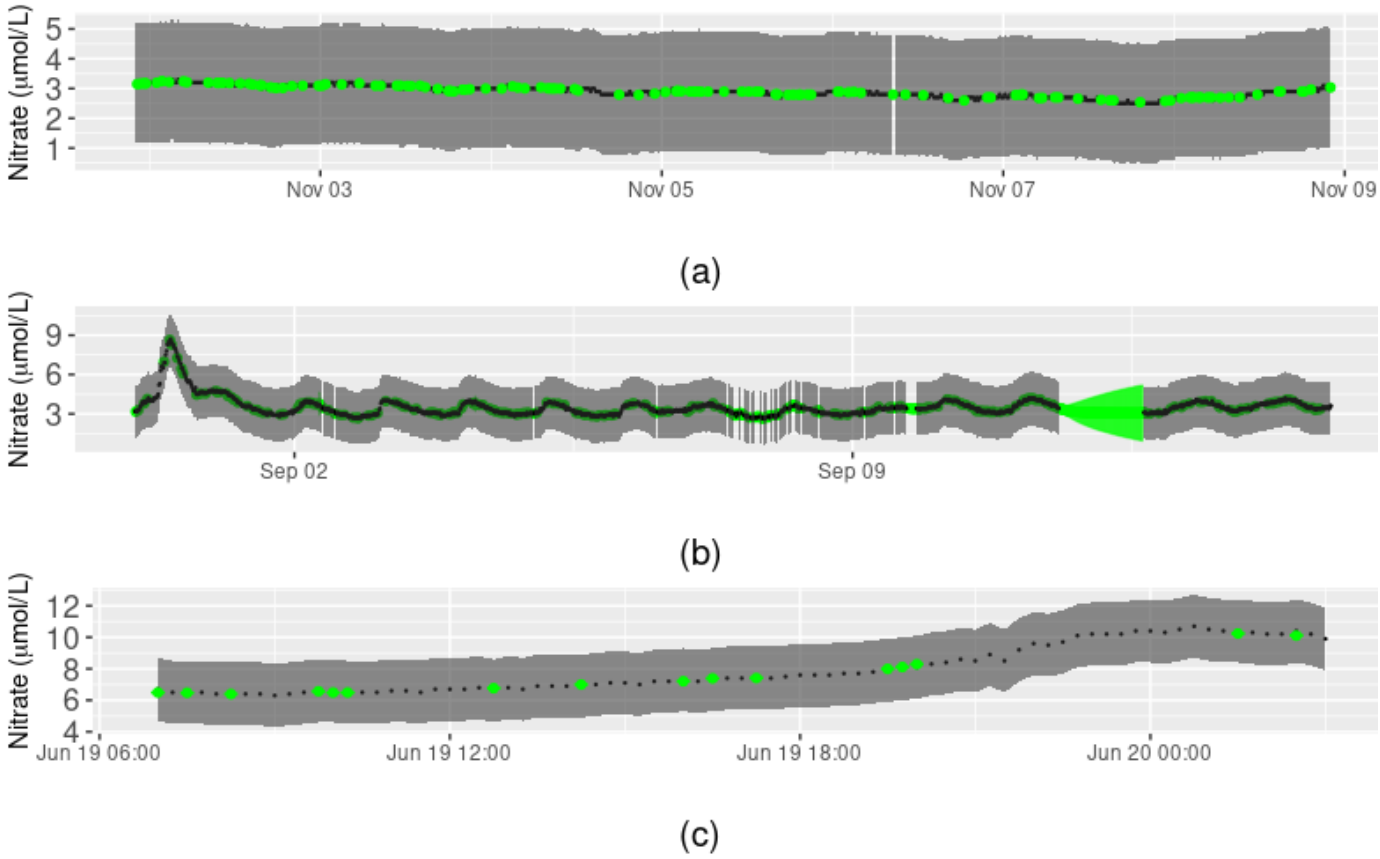

Figure 5. Performance evaluation: missing point data examples. Examples for different periods of randomly removed point observations, simulated 100 times: (a) one week, (b) one month, and (c) a nitrate event in which concentrations rose rapidly in less than 24 h. Dark points represent the real nitrate concentration value and the grey shading around those points the precision interval of the sensor. Green points and shading are the predicted values along with the prediction interval. Prediction intervals may not be visible when they are narrow relative to the precision interval of the sensor.

3.2.2.Simulations 4 and 5: missing sequences of data

When reconstructing missing sequences of data in the simulated timeseries, performance declined as sequence duration increased (i.e., the RMSE increased and PWPI decreased (Figure 6). The median and third quartile of RMSE for simulations where 10 day-long sequences of data were randomly removed were 0.25µmol/L and 0.44µmol/L, respectively, compared with 0.75µmol/L and 1.16µmol/L, respectively for simulations where 10 week-long sequences were randomly removed. For the median and third quartile PWPI, the one-day vs one-week comparisons were 0,7 and 0,74 vs 0,66 and 0,7 respectively.

We also observed that performance depended on whether GAM and ARIMA, GAM alone or ARIMA alone was used for the reconstruction. When ARIMA was used, the amount of missing data present in the preceding period also impacted performance. For example, both GAM and ARIMA were required for a week-long reconstruction in early March 2019 (Figure 7a), but this week occurred just after a three-month period of missing data such that the ARIMA (based on the previous 500 observations) was unable to perform well. The ARIMA always predicted a nitrate concentration of 4.5 µmol/L for missing data in the week-long sequence, whereas the GAM predictions followed the actual concentrations closely. This was also the case when GAM was used alone due to all covariates being available throughout the week-long sequence (Figure 7c).

In the case where ARIMA is used after a period with little to no missing values (Figure 7b), almost all real values of nitrate concentration were within the prediction intervals of the reconstructed data. However, the nitrate prediction interval increased as the number of timestamps into the future increased.

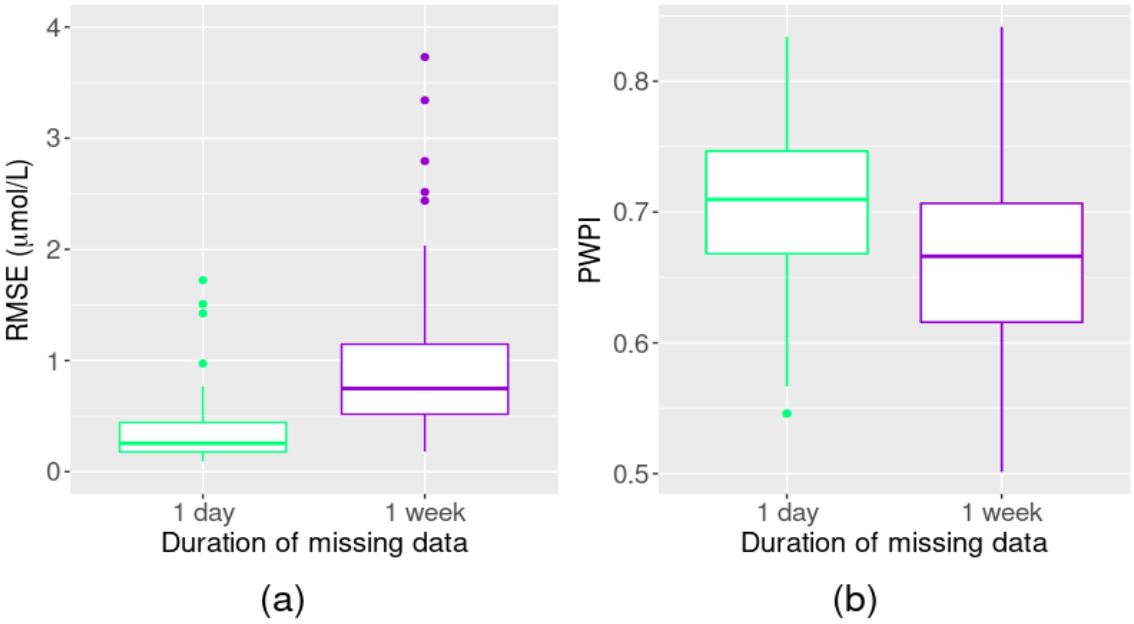

Figure 6: Performance evaluation: reconstructing sequences of missing data. Boxplots of (a) root-mean-squared-error (RMSE) and (b) the proportion of reconstructed data within the precision interval (PWPI) for different amounts of randomly removed point observations, simulated 100 times. Note that the y-axis on plot (a) has been truncated at 4 µmol/L (one extreme RMSE value of 45.23 µmol/L for the one-week simulations not shown).

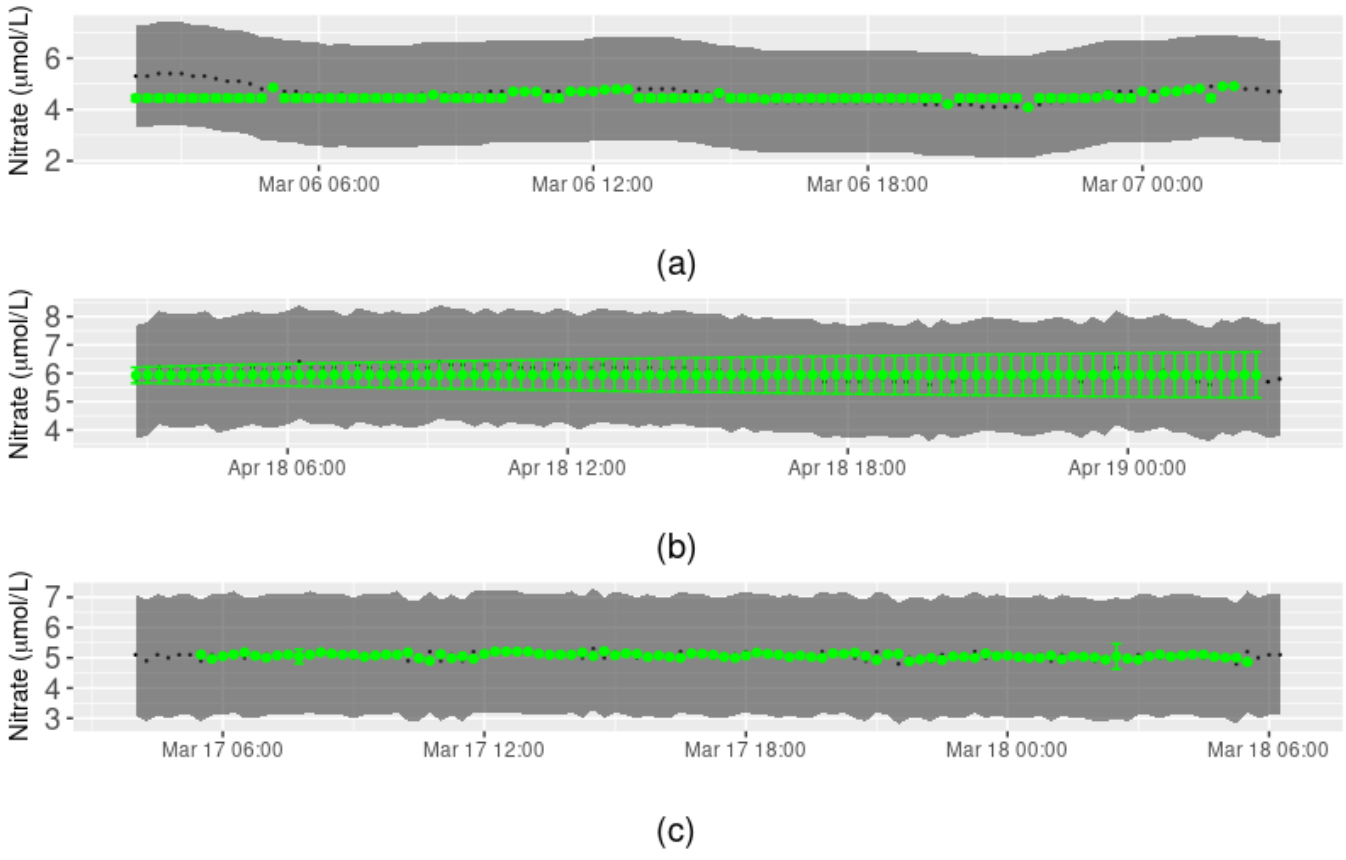

Figure 7: Performance evaluation: missing sequential data examples. Examples for different periods of randomly removed one-week sequences of observations, simulated 100 times, where data were reconstructed using (a) GAM and ARIMA, (b) ARIMA only, and (c) GAM only. Dark points represent the real nitrate concentration value and the shaded area around those points is the

precision interval of the sensor. Green points and shading are the predicted values along with the prediction interval. Prediction intervals may not be visible when they are narrow relative to the precision interval of the sensor.

4. Discussion

Data from low-cost, in-situ water-quality sensors provide unprecedented opportunities to better understand spatial and temporal water-quality dynamics. However, in-situ sensors are prone to technical issues, which presents a challenge for the processing and analysis of environmental data. The study presented here demonstrates that it is possible to predict these missing data for reconstruction of high-frequency environmental time series using appropriate statistical methods. To our knowledge, our study is among the first to reconstruct missing nitrate data from high-frequency data collected by in-situ sensors. This may be in part due to the relatively recent, standard use of such sensors for measuring nitrate concentration in river networks and makes comparison of our findings with other studies and methods difficult. Reconstruction of high-frequency runoff data showed that a new machine learning method, "nu-support vector machines," out-performed other machine learning methods [57]. Performance of the new method was evaluated in terms of the correlation (R^2) between observed and simulated data, with their method achieving values between 0.75 and 0.95. Applying the same R^2 coefficient to our simulations achieves R^2 values between 0.976 and 0,997 when 40% of the dataset is removed, and between 0.119 and 0.994 (first quartile = 0.84) when sequences of 10 days are removed, indicating that our method attains a comparatively good performance, particularly for point and short periods of missing data. Blending ARIMA forecasts and backcasts has also shown promise for reconstruction of sensor-based water-quality data including temperature, pH, specific conductance and dissolved oxygen [58]. However, we are unable to compare our results with this work given performance of the correction method was assessed by comparing the ARIMA-based results with corrections done manually by technicians.

Our method was able to predict all missing values present in a real time series of nitrate concentration data. Although the prediction interval for some predicted values were relatively wide, the median prediction interval was very low (1.34 μmol/L nitrate) indicating that many missing data had a 95% prediction interval lower than the sensor accuracy (i.e., at least 2 μmol/L) and therefore precise enough for the intended use of the data. We also showed via simulation that even when 40% of the initial dataset (point observations) were missing, our method was able to accurately recover approximately 70% of the data. When day-long sequences of contiguously missing data were simulated, mimicking for example a persistent sensor outage or prolonged periods of quality-flagged data, performance of the method was similarly efficient. However, for week-long periods of missing data, the percentage of accurately covered data decreased, indicating that data reconstruction is more impacted by long sequences of missing data in a row than by multiple but sporadically missing data.

Consideration of different periods of the nitrate concentration time series provided insight into the overall utility of the method and why the method may not accurately reconstruct all missing data. For example, excessive nitrate can create eutrophication issues in aquatic systems and therefore, for the purposes of environmental management, it is important to know (i) whether the presence of high nitrate concentrations is real or anomalous, and (ii) that accurate reconstruction of the real concentrations can be achieved in a timely fashion, particularly during floods. This appears possible with the method we have developed, given that missing values during periods of sudden rise and fall in nitrate concentration were predicted accurately (Figure 3b, Figure 5c). However, when environmental covariates from co-located sensors were not available, then reconstruction relied on ARIMA, for which prediction performance was inferior to that of GAM. This finding indicates the importance of having high-frequency sensors that can collect other environmental and water-quality data besides nitrate concentration at collection sites and is in accordance with results from a study of daily streamflow data [25] that found

increased accuracy of missing-data reconstruction when multiple input variables were included.

The method presented here was developed with the objective of being able to reconstruct data that are missing, for example due to their removal after being determined as technical anomalies, from environmental data collected by high-frequency in-situ sensors, using nitrate concentration data from Arikaree River. The method is currently applied in a binary fashion depending on the existence of covariates (other environmental data that can be used as predictors). For any one missing nitrate observation, GAM is used to predict the observation when data for all covariates are available, and ARIMA is used when data for at least one of the covariates is missing. Several avenues of study are envisaged from this work. First, future work could aim to develop a method whereby one or more environmental variables could be used as the covariate(s) according to their availability such that ARIMA was only used when no covariate data was available. A second avenue would be to use different types of models in the framework, bearing in mind that the characteristics of time series data can influence the forecasting method that should be run. Other methods that may be suitable for the particular characteristics of water-quality time series include Seasonal Autoregressive Integrated Moving Average (SARIMA) for seasonal data or deep learning methods as Long Short-Term Memory (LSTM) networks. Finally, future research could seek to confirm the applicability of our method to other sites and environmental data in order to generalize the framework.

5. Conclusions

Measurement errors or missing observations are recurrent and, in some cases, may reduce user perception of data quality, thereby preventing data from underpinning management actions. Here, we developed a method to successfully reconstruct missing nitrate-concentration data from high-frequency in-situ sensors in fresh waters, thereby adding value to the literature on anomaly detection and filing a critical management need in the environmental domain. To mimic sporadically missing observations, both point data and sequences of data were removed from a two-year time series of nitrate-concentration data. In 72% of cases with missing point data, predicted values were within the sensor-precision interval of the original value, although the predictive ability declined when sequences of missing data occurred. The models also had stronger predictive ability when other water variables (covariates) were available. This suggests there may be advantages to deploying co-located sensors to measure covariates, even when there is a single constituent of concern such as nitrate by enabling a more reliable reconstruction of the nitrate time series. Our study is an important first step towards environmental data reconstruction in the information age and sets a benchmark against which future datasets and methodological developments can be compared. While we believe the general methodology presented here is generalizable to rivers in other ecosystems [59], the relationships between other water-quality variables of interest may differ. Thus, future research should also focus on understanding these relationships so that co-located sensors can be optimally deployed. This will ensure that near real-time water-quality data produced by low-cost in-situ sensors are trustworthy and reliable enough to underpin data-enabled management decisions.

Author Contributions: Conceptualization, E.P. and K.M.; methodology, C.K., R.H., K.M., and B.L.; formal analysis, C.K. and R.H.; resources, G.L. and J.J.; writing—original draft preparation, C.K., E.P., G.L., J.J and C.L.; writing—review and editing, all authors; visualization, C.K. and C.L.; supervision, C.L., E.P. and R.H.; funding acquisition, B.L. and K.M. All authors have read and agreed to the published version of the manuscript.

Funding: Funding was provided by the Energy Environment Solutions (E2S-UPPA) consortium through an international research chair. This study was part of a project funded by an Australian Research Council (ARC) Linkage grant (LP180101151) "Revolutionising water-quality monitoring in the information age."

Data Availability Statement: Data are available through the NEON website https://www.neonscience.org/. The National Ecological Observatory Network is a program sponsored by the National Science Foundation and operated under cooperative agreement by Battelle Memorial Institute. This material is based in part upon work supported by the National Science Foundation through the NEON Program.

Acknowledgments: The authors acknowledge and thank the Queensland Department of Environment and Science, and in particular the Water Quality and Investigations team for their valuable discussions regarding the ARC Linkage project LP180101151. We extend our thanks to all those involved across project.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. NEON data. Details on sensors, variables collected, units of measurement, associated data-collection intervals, and the NEON data product number for data used in this study.

Sensor	Water-quality variable	Unit	Published data resolution	Published interval (min)	Product number
SUNA v2	Nitrate	μmol NO_3-N /L	0.1	15	DP1.20033.001
YSI EXO Optical Dissolved Oxygen	Dissolved oxygen	mg/L	0.01	1	DP1.20288.001
Level TROLL 500	Water elevation	masl	0.01	5	DP1.20016.001
YSI EXO Turbidity	Turbidity	FNU	0.01	1	DP1.20288.001
YSI EXO Conductivity and Temperature	Specific conductance	μS/cm	0.01	1	DP1.20288.001
Platinum Resistance Thermometer	Water temperature	°C	0.01	1	DP1.20053.001

References

1. K. S. Adu-Manu, C. Tapparello, W. Heinzelman, F. A. Katsriku, J.-D. Abdulai, Water quality monitoring using wireless sensor networks: Current trends and future research directions, ACM Transactions on Sensor Networks (TOSN) 13 (1) (2017) 1–41. doi:10.1145/3005719
2. F. A. Katsriku, M. Wilson, G. G. Yamoah, J.-D. Abdulai, B. M. A. Rahman, K. T. V. Grattan, Framework for Time Relevant Water Monitoring System, Springer International Publishing, Cham, 2015, pp. 3–19. doi:10.1007/978-3-319-08239-4_1
3. A. S. Jones, J. S. Horsburgh, S. L. Reeder, M. Ramirez, J. Caraballo, A data management and publication workflow for a large-scale, heterogeneous sensor network, Environmental Monitoring and Assessment 187 (6) (2015) 348. doi:10.1007/s10661-015-4594-3
4. J. Park, K. T. Kim, W. H. Lee, Recent advances in information and communications technology (ICT) and sensor technology for monitoring water quality, Water 12 (2) (2020) 510. doi:10.3390/w12020510
5. K. M. Cawley, Neon algorithm theoretical basis document (atbd), Tech. rep. (2018).
6. C. Leigh, O. Alsibai, R. J. Hyndman, S. Kandanaarachchi, O. C. King, J. M. McGree, C. Neelamraju, J. Strauss, P. D. Talagala, R. D. Turner, K. Mengersen, E. E. Peterson, A framework for automated anomaly detection in high frequency water-quality data from in situ sensors, Science of the Total Environment 664 (2019) 885 – 898. doi:10.1016/j.scitotenv.2019.02.085
7. C. Leigh, O. Alsibai, R. J. Hyndman, S. Kandanaarachchi, O. C. King, J. M. McGree, C. Neelamraju, J. Strauss, P. D. Talagala, R. D. Turner, et al., A framework for automated anomaly detection in high frequency water-quality data from in situ sensors, Science of the Total Environment 664 (2019) 885–898. doi:10.1016/j.scitotenv.2019.02.085
8. J. Rodriguez-Perez, C. Leigh, B. Liquet, C. Kermorvant, E. Peter son, D. Sous, K. Mengersen, Detecting technical anomalies in high frequency water-quality data using artificial neural networks, Environmental Science & Technology 54 (21) (2020) 13719–13730. doi:10.1021/acs.est.0c04069

9. J. Liu, P. Wang, D. Jiang, J. Nan, W. Zhu, An integrated data-driven framework for surface water quality anomaly detection and early warning, Journal of Cleaner Production 251 (2020) 119145. doi:10.1016/ j.jclepro.2019.119145
10. B. Shi, P. Wang, J. Jiang, R. Liu, Applying high-frequency surrogate measurements and a wavelet-ANN model to provide early warnings of rapid surface water quality anomalies, Science of The Total Environment 610-611 (2018) 1390–1399. doi:10.1016/j.scitotenv.2017.08.232
11. G. E. Batista, M. C. Monard, An analysis of four missing data treatment methods for supervised learning, Applied Artificial Intelligence 17 (5-6) (2003) 519–533. doi:10.1080/713827181
12. Y. Dong, C.-Y. J. Peng, Principled missing data methods for researchers, SpringerPlus 2 (1) (2013) 1–17. doi:10.1186/2193-1801-2-222
13. J. Hannaford, G. Buys, Trends in seasonal river flow regimes in the UK, Journal of Hydrology 475 (2012) 158–174. doi: 10.1016/j.jhydrol.2012.09.044
14. D. R. Helsel, R. M. Hirsch, Statistical Methods in Water Resources, Vol. 49, Elsevier, 1992.
15. L. Slater, G. Villarini, On the impact of gaps on trend detection in extreme streamflow time series, International Journal of Climatology 37 (10) (2017) 3976–3983. doi:10.1002/joc.4954
16. R. M. Hirsch, An evaluation of some record reconstruction techniques, Water Resources Research 15 (6) (1979) 1781–1790. doi:10.1029/ WR015i006p01781
17. J. R. Wallis, D. P. Lettenmaier, E. F. Wood, A daily hydroclimatological data set for the continental United States, Water Resources Research 27 (7) (1991) 1657–1663. doi:10.1029/91WR00977
18. H. Raman, S. Mohan, P. Padalinathan, Models for extending streamflow data: a case study, Hydrological Sciences Journal 40 (3) (1995) 381–393. doi:10.1080/02626669509491422
19. C. A. Woodhouse, S. T. Gray, D. M. Meko, Updated streamflow reconstructions for the Upper Colorado River Basin, Water Resources Research 42 (5) (2006). doi:10.1029/2005WR004455
20. B. A. Amisigo, N. C. van de Giesen, Using a spatio-temporal dynamic state-space model with the EM algorithm to patch gaps in daily riverflow series, Hydrology and Earth System Sciences 9 (3) (2005) 209–224. doi: 10.5194/hess-9-209-2005.
21. M. Khalil, U. Panu, W. Lennox, Groups and neural networks based streamflow data infilling procedures, Journal of Hydrology 241 (3) (2001) 153–176. doi:10.1016/S0022-1694(00)00332-2
22. A. Elshorbagy, S. Simonovic, U. Panu, Estimation of missing streamflow data using principles of chaos theory, Journal of Hydrology 255 (1) (2002) 123–133. doi:10.1016/S0022-1694(01)00513-3
23. P. Coulibaly, C. K. Baldwin, Nonstationary hydrological time series forecasting using nonlinear dynamic methods, Journal of Hydrology 307 (1) (2005) 164–174. doi:10.1016/j.jhydrol.2004.10.008
24. C. L. Harvey, H. Dixon, J. Hannaford, An appraisal of the performance of data-infilling methods for application to daily mean river flow records in the UK, Hydrology Research 43 (5) (2012) 618–636. doi:10.2166/ nh.2012.110
25. P. Tencaliec, A.-C. Favre, C. Prieur, T. Mathevet, Reconstruction of missing daily streamflow data using dynamic regression models, Water Resources Research 51 (12) (2015) 9447–9463. doi:10.1002/ 2015WR017399
26. L. Zhao, F. Zheng, Missing data reconstruction using adaptively updated dictionary in wireless sensor networks, in: Proceedings of the 2017 The 7th International Conference on Computer Engineering and Networks (2017) 22–23.
27. P. Liqiang, L. Jianzhong, A multiple-regression-model-based missing values imputation algorithm in wireless sensor network, Journal of Computer Research and Development 46 (12) (2009) 2101.
28. H. Wu, J. Xian, J. Wang, S. Khandge, P. Mohapatra, Missing data recovery using reconstruction in ocean wireless sensor networks, Computer Communications 132 (2018) 1–9. doi:10.1016/j.comcom.2018. 09.007
29. C.-M. Lee, C.-N. Ko, Time series prediction using RBF neural networks with a nonlinear time-varying evolution PSO algorithm, Neurocomputing 73 (1) (2009) 449–460. doi:10.1016/j.neucom.2009.07.005
30. S. Jeong, M. Ferguson, R. Hou, J. P. Lynch, H. Sohn, K. H. Law, Sensor data reconstruction using bidirectional recurrent neural network with application to bridge monitoring, Advanced Engineering Informatics 42 (2019) 100991. doi:10.1016/j.aei.2019.100991
31. J. A. Camargo, A. Alonso, A. Salamanca, Nitrate toxicity to aquatic animals: a review with new data for freshwater invertebrates, Chemosphere 58 (9) (2005) 1255–1267. doi:10.1016/j.chemosphere.2004.10.044
32. D. F. Boesch, W. R. Boynton, L. B. Crowder, R. J. Diaz, R. W. Howarth, L. D. Mee, S. W. Nixon, N. N. Rabalais, R. Rosenberg, J. G. Sanders, et al., Nutrient enrichment drives Gulf of Mexico hypoxia, EOS, Transactions American Geophysical Union 14 (2009) 117–118. doi:10.1029/2009EO140001
33. S. B. Bricker, B. Longstaff, W. Dennison, A. Jones, K. Boicourt, C. Wicks, J. Woerner, Effects of nutrient enrichment in the nation's estuaries: a decade of change, Harmful Algae 8 (1) (2008) 21–32. doi:10.1016/j.hal.2008.08.028
34. J. A. Camargo, J. Ward, Nitrate (NO3-N) toxicity to aquatic life: A proposal of safe concentrations for two species of Nearctic freshwater invertebrates, Chemosphere 31 (5) (1995) 3211–3216. doi:10.1016/ 0045-6535(95)00182-8
35. J. Davidson, C. Good, C. Williams, S. T. Summerfelt, Evaluating the chronic effects of nitrate on the health and performance of postsmolt Atlantic salmon *Salmo salar* in freshwater recirculation aquaculture systems, Aquacultural Engineering 79 (2017) 1 – 8. doi: 10.1016/j.aquaeng.2017.08.003
36. A. P. Moore, R. B. Bringolf, Effects of nitrate on freshwater mussel glochidia attachment and metamorphosis success to the juvenile stage, Environmental Pollution 242 (2018) 807–813. doi:10.1016/j.envpol. 2018.07.047

37. C. Leigh, M. A. Burford, R. M., Connolly, J. M., Olley, E. Saeck, F. Sheldon, J. C. Smart, S. E. Bunn, Science to support management of receiving waters in an event-driven ecosystem: From land to river to sea. Water 5 (2) (2013) 780-797. doi:10.3390/w5020780
38. K. R. O'Brien, T. R. Weber, C. Leigh, M. A. Burford, Sediment and nutrient budgets are inherently dynamic: evidence from a long-term study of two subtropical reservoirs. Hydrology and Earth System Sciences 20 (12) (2016) 4881-4894. doi:10.5194/hess-20-4881-2016
39. S.G Fisher, N.B. Grimm, E. Martí, R. M. Holmes, J. B. Jones Jr, Material spiraling in stream corridors: a telescoping ecosystem model. Ecosystems 1 (1) (1998) 19-34. doi:10.1007/s100219900003
40. S. N. Wood, Generalized additive models: An introduction with R, CRC press, 2017.
41. Y. Sakamoto, M. Ishiguro, G. Kitagawa, Akaike information criterion statistics, Dordrecht, The Netherlands: D. Reidel 81 (1986). doi:10. 1080/01621459.1988.10478680
42. G. E. Box, D. A. Pierce, Distribution of residual autocorrelations in autoregressive-integrated moving average time series models, Journal of the American Statistical Association 65 (332) (1970) 1509–1526. doi:10.1080/01621459.1970.10481180
43. J. Vance, B. Nance, D. Monahan, M. Mahal, M. Cavileer, Neon preventive maintenance procedure: AIS surface water quality multisonde. neon.doc.001569 revision: B, Tech. rep., National Ecological Observatory Network (2019). URL http://data.neonscience.org
44. R. Willingham, M. Cavileer, J. Csavina, D. Monahan, Neon preventive maintenance procedure: Submersible ultraviolet nitrate analyzer. neon.doc.002716 revision: B, Tech. rep., National Ecological Observatory Network (2019). URL http://data.neonscience.org
45. K. M. Cawley, Neon aquatic sampling strategy, Tech. rep., National Ecological Observatory Network (2016). URL http://data.neonscience.org
46. NEON, Nitrate in surface water (DP1.20033.001) (2021). doi:10.48443/924T-1K41. URL https://data.neonscience.org/data-products/DP1.20033.001/RELEASE-2021
47. NEON, Water quality (DP1.20288.001) (2021). doi:10.48443/ D8KW-5J62. URL https://data.neonscience.org/data-products/DP1.20288.001/RELEASE-2021
48. NEON, Temperature (PRT) in surface water (DP1.20053.001) (2021). doi:10.48443/NY19-PJ91. URL https://data.neonscience.org/data-products/DP1.20053.001/RELEASE-2021
49. NEON, Elevation of surface water (DP1.20016.001) (2021). doi:10.48443/QSER-8M94. URL https://data.neonscience.org/data-products/DP1.20016.001/RELEASE-2021
50. C. Leigh, S. Kandanaarachchi, J.M. McGree, R.J. Hyndman, O. Alsibai, K. Mengersen, K., E.E. Peterson, Predicting sediment and nutrient concentrations from high-frequency water-quality data. PloS One 14 (8) (2019) e0215503. doi:10.1371/journal.pone.0215503
51. G. E. Box, D. R. Cox, An analysis of transformations. Journal of the Royal Statistical Society: Series B (Methodological) 26 (2) (1964) 211-243. doi:10.1111/j.2517-6161.1964.tb00553.x
52. R Core Team, R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing (2020). URL https://www. R-project. org/
53. J. Fox, S. Weisberg, D. Adler, D. Bates, G. Baud-Bovy, S. Ellison, D.Firth, M. Friendly, G. Gorjanc, S. Graves, et al., Package 'car', Vienna: R Foundation for Statistical Computing (2012).
54. T. Hastie, gam: Generalized Additive Models, R package version 1.20 (2020). URL https://CRAN.R-project.org/package=gam
55. R. J. Hyndman, G. Athanasopoulos, C. Bergmeir, G. Caceres, L. Chhay, M. O'Hara-Wild, M., F. Petropoulos, S. Razbash, E. Wang, Package 'forecast'. (2020). URL: https://cran. r-project. org/web/packages/forecast/forecast. pdf.
56. J. Taylor, S. Street, C. Sturtevant, Neon algorithm theoretical basis document (atbd): QA/QC plausibility testing. neon.doc.011081 revision: C, Tech. rep., National Ecological Observatory Network (2020). URL http://data.neonscience.org
57. J. Langhammer, J. Česák, Applicability of a nu-support vector regression model for the completion of missing data in hydrological time series. Water 8 (12) (2016) 560. doi:10.3390/w8120560
58. A. S. Jones, T. L. Jones, J. S. Horsburgh, Toward automating post processing of aquatic sensor data. Earth Archiv preprint (2021). doi: 10.31223/X5Z62X
59. C. Kermorvant, B. Liquet, G. Litt, K. Mengersen, E.E. Peterson, R. Hyndman, J.B. Jones Jr, C. Leigh, Understanding links between water-quality variables and nitrate concentration in freshwater streams using high-frequency sensor data. (2021) arXiv preprint arXiv:2106.01719.