Detecting technical anomalies in high-frequency water-quality data using Artificial Neural Networks

Javier Rodriguez Perez, Catherine Leigh, Benoit Liquet, Claire Kermorvant, Erin Peterson, Damien Sous, Kerrie L. Mengersen

Detecting technical anomalies in high-frequency water-quality data using Artificial Neural Networks

Javier Rodriguez-Perez,[†] Catherine Leigh,[¶,§,‖] Benoit Liquet,[†,⊥] Claire Kermorvant,[†,§] Erin Peterson,[§,‖,#] Damien Sous,[†,@] and Kerrie Mengersen[*,†,§,#]

[†]*Université de Pau et des Pays de l'Adour, E2S UPPA, CNRS, Pau, France*

[‡]*Institute for Multidisciplinary Research in Applied Biology, Departamento Ciencias del Medio Natural, UPNA Pamplona, Spain*

[¶]*Biosciences and Food Technology Discipline, School of Science, RMIT University,*
1 *Bundoora, Australia*

[§]*Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS), Australia*

[‖]*Institute for Future Environments, Queensland University of Technology, Brisbane, Australia*

[⊥]*Department of Mathematics and Statistics, Macquarie University, Sydney, Australia*

[#]*School of Mathematical Sciences, Queensland University of Technology, Brisbane, Australia*

[@]*Université de Toulon, Aix Marseille Université, CNRS, IRD, Mediterranean Institute of Oceanography (MIO), La Garde, France*

E-mail: k.mengersen@qut.edu.au

2 **Abstract**

Anomaly detection (AD) in high-volume environmental data requires one to tackle a series of challenges associated with the typical low frequency of anomalous events, the broad-range of possible anomaly types and local non-stationary environmental conditions, suggesting the need for flexible statistical methods that are able to cope with unbalanced high-volume data problems. Here, we aimed to detect anomalies caused by technical errors in water-quality (turbidity and conductivity) data collected by automated *in-situ* sensors deployed in contrasting riverine and estuarine environments. We first applied a range of Artificial Neural Networks (ANN) that differed in both learning method and hyper-parameter values, then calibrated models using a Bayesian multi-objective optimisation procedure, and selected and evaluated the "best" model for each water-quality variable, environment and anomaly type. We found that semi-supervised classification was better able to detect sudden spikes, sudden shifts and small sudden spikes whereas supervised classification had higher accuracy for predicting long-term anomalies associated with drifts and periods of otherwise unexplained high-variability.

Introduction

Monitoring the water quality of rivers is becoming increasingly relevant in the Anthropocene as the need for management strategies to safeguard water resources in human dominated landscapes accelerates. With the advent of new technologies, data on water properties can be obtained with high frequency in near-real time, which can facilitate fast and adaptive management strategies.[1,2] However, the generation of high-volume, high-velocity data can create problems of quality control due to a combination of (1) technical issues with the water-quality sensors that result in technical anomalies, and (2) operational problems with traditional, often manual and periodic, methods of quality control that are no longer feasible. In order to improve the quality and utility of high-volumes of data, we therefore need to explore efficient methods to complement or replace traditional methods of quality control, including manual anomaly detection and correction and quality coding.[3,4] Improving water-

quality monitoring will also require robust and flexible statistical methods that can handle data streaming, while also providing big data tools for data-driven decision making.[5–7]

The accurate detection of anomalies in high-frequency, high-volume water-quality data is additionally challenged by several factors. For example, the common low frequency of anomalies in water-quality time series constrains the ability to not only detect anomalies but also to evaluate the performance of detection methods.[8,9] There is also a broad range of possible anomaly types, ranging from sudden changes in water-quality values to anomalously constant values maintained for long periods.[10,11] In addition, the differentiation of technical anomalies, e.g. related to sensor malfunction or database failure, from the real deviation of the natural system from its typical behaviour is not always straightforward.[1] This is because natural fluctuations in environmental conditions at a given site can influence the level of detectability of sensor-related anomalies from the normal signal: the greater the fluctuation, the harder the detection of technical errors. All of these issues can constrain the generality and robustness of detection methods under a broad set of environmental conditions.[10,12]

A combination of methods may be required to address the challenges raised above. For instance, sudden spikes and shifts in values and impossible and/or out-of-sensor-range values on single or just a few observation points can be easily differentiated from adjacent, non-anomalous values. Despite the extremely low frequency of such point-based anomalies in time-series data, rule-based, regression-based and feature-based methods may provide optimal performance in terms of their detection.[1,11] Common anomalies in chemical and biochemical data from *in-situ* sensors also comprise multiple observations, such as sensor drift and periods of unusually high or low variability, which may indicate the need for sensor calibration or maintenance.[13] The detection of these multiple-points anomalies can be tackled by various methods,[7,11] but remains challenging and often still requires user intervention. Several methods can be combined to extend the range of detected anomalies. Recent work by Leigh et al.[11] showed that a combination of rule-, feature and regression-based methods facilitated the correct classification of impossible values, sudden isolated spikes and level

shifts, although drift and periods of high variability still tended to be associated with high rates of false positives.

Considering the current global effort to produce real-time, high-frequency and long-term monitoring of aquatic ecosystems, and the limitations of existing quality control approaches,[1,4,13] there is a pressing need to explore new methods. Artificial Neural Networks (ANNs) are a promising alternative.[6,7,14] One basic strength of ANNs in analysing high-frequency time-series data is that they do not require *a priori* knowledge of the underlying physical and environmental processes. ANNs have therefore been used tackle multiple problems of data streaming including predicting temporal patterns in both marine[15,16] and freshwater systems.[6,17,18] They have potential to detect anomalies and sudden changes that occur over short-time intervals in water-quality data,[6,18,19] suggesting that they can be readily applied to optimise anomaly detection in complex and fluctuating environmental systems. Previous work on AD applied to water-quality variables compared ANNs to other machine learning methods and found that ANNs have comparable performance to other methods in terms of AD.[12,18,20,21] However, authors also acknowledge that remains understudied the optimisation techniques (especially on those machine learning methods whose learning process depends on a large number of hyper-parameters) that aim to improve detection capacity under real-world conditions.[7,20]

The aim of this study is to test the ability of ANNs to detect technical anomalies in water-quality time-series data in contrasting riverine and estuarine environments. We were especially interested in testing the ability of ANNs to detect multiple-point anomalies, for example those that may result from sensor drift, given that methods previously examined were limited in their ability to accurately detect such context-dependent events. The studied variables are turbidity and conductivity collected at high-frequency by autonomous *in-situ* sensors. The monitored sites in subtropical Australia and temperate France provide contrasting ranges of environmental conditions, particularly in terms of the studied variables and their magnitudes and dynamics. The comparison and evaluation of detection methods

in regions that vary in environmental conditions and across a common range of variables is essential to determine the methods' broad suitability for anomaly detection.[7] Given that the environmental fluctuations in each river (freshwater and estuarine sites) may influence the ability to detect the different types of anomalies, we calibrated ANNs using models that differed in learning method and hyper-parameter values. To do so, we implemented a Bayesian optimisation method using a multi-objective approach (i.e. based on scores specially suited to unbalanced classification problems), in order to find the combination of ANN hyper-parameters best suited for anomaly detection at each location. The performance of the calibrated ANNs in detecting different types of anomalies was also evaluated and compared to regression-based methods developed for and applied to similar data.[11]

Materials and Methods

Study sites

The studied water-quality data derive from two freshwater rivers in north eastern Australia, the Pioneer River and Sandy Creek, and an estuarine river in south western France, the Ardour River estuary. Both study areas have seasonality in climate and therefore provide non-stationary environmental conditions of water properties throughout the year.

The Australian study area is characterised by humid subtropical climate and strong seasonality: the wet season (typically occurring between December and April) has higher rainfall and air temperatures than the dry season which has lower rainfall and is associated with low to zero river flows.[22] Pioneer River (PR) is in the Mackay Whitsunday region of northeast Australia, with a length of 120 km and a monitored catchment area of 1,466 km^2, with the upper reaches flowing predominantly through National or State Parks and its middle and lower reaches flowing through land dominated by sugarcane farming. Sandy Creek (SC) is a low-lying coastal-plain stream, 72 km long with a monitored catchment area of 326 km^2 and with a similar land-use and land-cover profile to that of the lower Pioneer River. Both

study sites are in the freshwater reaches of these rivers.

In contrast, south-western France is characterised by an oceanic temperate climate, with drier and warmer conditions from April to October. The Adour River (AR) in the Aquitaine region of southwestern France is 330 km in length and has a catchment area of 16,880 km^2, with reaches flowing throughout a mosaic of agricultural and forested lands. The study site is in the lower estuary between Bayonne city and the river mouth about 3 km downstream. The tidal range varies from 1 to 4.5 m. Under the combined influence of rainfall and snowmelt during late spring, river flow is highly variable, with minima and maxima of about 80 and 3000 m^3/s, respectively.[23] Driven by such strong fluctuations in tidal range and river discharge, the estuary is classified as a highly variable, time-dependent salt-wedge.[23]

Water-quality data

Autonomous multi-parameter *in-situ* sensors have been installed at each site. At PR and SC, YSI EXO2 sondes with YSI Smart Sensors are housed in flow cells in water-quality monitoring stations on the rivers' banks. At pre-defined time intervals (see bellow), monitoring systems transport water via a pumping system from the river up to the flow cell. Sensors are equipped with wipers to minimise biofouling and all equipment undergo regular maintenance and calibration, with sensors calibrated and equipment checked approximately every six weeks following manufacturer guidelines. At AR, a YSI 6920 sonde is installed 1 m below the water surface on a floating pier, providing direct measurement of water quality in the river. Data are retrieved during maintenance operations (sensor cleaning, memory erasing, battery replacement) approximately every eight weeks.

At each site, the *in-situ* sensors record turbidity (NTU) and electrical conductivity (conductivity; μS/cm). The temporal resolution of data differs among rivers according to local variability in environmental conditions; measurements are taken every 60 and 90 minutes in PR and SC, respectively, and every 10 minutes in the AR. At PR and SC, measurements are sometimes also taken at more frequent intervals during high flow events. The data se-

quences are one year (12 March 2017 to 12 March 2018) and five months (11 March to 13 August 2019) for the Australian and French sites, respectively, totalling 6280, 5402 and 14,185 observations at PR, SC and AR, respectively.

Water-quality parameters were strongly affected by the non-stationary flow regimes of each site (see Fig. 1). Turbidity is a visual property of water clarity with higher values indicating lower clarity. High turbidity often occurs during flood events, when rivers become loaded with sediment eroded from the catchment, and during local re-suspension events that can result from salt-wedge arrival, flow instability, or navigational and dredging activities. In contrast, low values tend to occur during low discharge periods, e.g. during the dry season, when the slow currents promote sediment deposition or, in estuarine environments, when low-turbidity marine waters have entered the system. These overall trends can become much more complex, and sometimes reversed, under the influence of specific local features. Conductivity reflects the concentration of ions in the water. The dominant effect is generally due to salts, such as sodium chloride, with marine waters typically having much higher conductivity than inland waters. Other dissolved ionic compounds such as nutrients also contribute to conductivity.

In the Australian sites, the water regime is driven by rainfall patterns, generating strong positive or negative phases for either turbidity and conductivity values, coinciding with new and sudden inputs of fresh water or periods of low and zero flow (see upper and mid panels in Fig. 1). By contrast, AR is strongly conditioned by the interplay between river discharge and tide forcing. The basic trend is that during high-discharge events, the turbidity reaches high values corresponding to massive seaward fluxes of suspended sediments, and the conductivity is minimal because the marine waters are flushed out from the estuary (see lower panels in Fig. 1). During moderate discharge events, the tidal influence becomes dominant and drives tidal fluctuations of conductivity and turbidity inside the estuary: at high tide, the estuary is filled by salty (high conductivity) and clearer (low turbidity) marine waters while the reverse behaviour is observed at low tide. This pattern is further affected by strong vertical

gradients, the salty marine waters being heavier than fresh riverine waters, which induces complex time-dependent dynamics at the tidal scale.[23]

We focused on turbidity and conductivity data, two common water quality-variables measured in the majority of monitoring programs worldwide and commonly used as proxies of nutrient and/or sediment pollution in rivers, estuaries and marine environments.[24] As a global problem, sediment pollution also causes major economic issues related drinking water and dredging.[25] In addition, turbidity and conductivity are typically more stable in rivers through time than other properties such as dissolved oxygen and water temperature, which fluctuate daily as well as seasonally.

Definitions and types of anomalies

Environmental sensors generating high-frequency data sets are usually subject to technical anomalies that can derive from fouling of sensors, sensor calibration shifts, power supply problems or unforeseen environmental conditions that adversely affect the sensor equipment.[1,26] In general, such technical anomalies strongly depart from the expected pattern of the non-anomalous observations.[10] In addition, strong inter- and intra-seasonal variations in environmental conditions from site to site may create local differences in the occurrence and types of anomalies, which can constrain their transferability to other sites.[27] We thus need to develop methods for detecting these technical anomalies, that can overcome these challenges,[1,11,27] particularly in the context of data streaming.

The present analysis focuses on technical anomaly detection (AD) in water-quality time series collected by *in-situ* sensors. Types of such anomalies have been described[1,11] and previous work has indicated the need to explore new statistical methods to better detect the full suite of these types, notably those occurring over multiple, continuous observations (for example as a result of sensor drift), given that such anomalies commonly occur in water-quality time series yet are not as successfully detected as other types of anomalies (e.g. single-point anomalies, out-of-range values).[11]

Following the framework provided by,[11] a range of technical anomalies in the time-series was first identified by local water-quality experts (i.e. for SC and PR by[11] and for AR by DS). For each site and variable (i.e. turbidity and conductivity), the identified anomalies were labelled along with their types based on anomaly classes defined by.[11] In the present work, we first focused on anomaly types of Class 1 defined as a single observation generating sudden changes in value from the previous observation. Class 1 anomalies were also considered a high priority in terms of detection, given that (true) sudden changes in turbidity and conductivity may be used as early warning signals of water quality by local environmental agencies and that they can strongly influence water-quality assessments and consequent management decisions. We additionally focused on Class 3 anomalies, which include technical anomalies such as long-term calibration offsets and changes comprised of multiple dependent observations. Class 3 was considered lower priority than Class 1 and may require *a posteriori* user intervention (i.e. after data collection rather than in real time) to confirm observations as anomalous. Specifically, using the nomenclature introduced by,[11] we based our analyses of AD on those types of Class 1 anomalies defined as (a) large sudden spikes (type A), sudden shifts (type D), small sudden spikes (type J), and Class 3 anomalies defined as drift (type H), high variability (type E) and untrustworthy data not defined by other types (type L). For more information about definitions of each anomaly type and class see.[11]

The present analysis does not focus on AD of those types associated with impossible, out-of-sensor-range and missing values (Class 2) given that they can be easily detected by automated, hard-coded classification rules.[11] For each site, we therefore removed all Class 2 anomalies from the data prior to analysis. The filtered data were then log-transformed to remove exponential data variance (see Fig. 1) to produce the time-series used for the presented analysis.

Learning methods and data processing

One of the most challenging issues for AD in time-series data is that most data are usually "normal" or non-anomalous, while anomalous values are rare.[8,9] One of the most common discriminating approaches to solve such a problem is to compare the similarity between two sequences of time-dependent variables, specifically between the sequence of observed and predicted values.[11,28] When the anomalous values are pre-labelled, we can apply semi-supervised classification based on training the learning process with the non-anomalous data, fitting models with prediction errors and predicting anomalous events as those observed values falling outside prediction intervals.[11,28] Supervised classification feeds the learning process with both a sequence of labelled values for AD (including both anomalous and non-anomalous values, tagged accordingly). The learning process can then generate a sequence of probabilities which can be binary-classified as anomalous or non-anomalous according to a predefined threshold parameter.

For each water-quality variable and site, time-series data were partitioned according to each learning process (i.e. semi-supervised or supervised classification). For semi-supervised classification, we retained the "normal" values (i.e. we discarded the anomalous values) and divided the time-series into four contiguous sequences of equal size: namely, two adjacent sequences of values for training and two other adjacent sequences for validation (for details, see Figure S1 in SI). Each pair of contiguous sequences for training or validation comprised one sequence for prediction (predictor variable) followed by an adjacent sequence for the outcome (outcome variable), respectively. For supervised classification, the sequence of values (including both anomalous and "normal" values) for prediction and the sequence of labelled values for AD were each divided into two adjacent sequences of equal length, for training and validation (for details, see Figure S1 in SI). Thus, predictor and outcome variables were composed of one sequence for training followed by an adjacent sequence for validation, respectively. Given the proportionally low number of anomalies in our data set, we did not use a "test" data set during the optimisation process.

Water quality in each of the studied sites undergoes intra- and interannual variation associated with local, seasonal and annual cycles and stochastic environmental events. We therefore incorporated within our analysis a "sliding window" or moving sequence of values of a defined length along the time series. For each site, the sliding window was defined according to the temporal resolution of data, and was 60, 90 and 10 minutes for PR, SC and AR, respectively (see above). For both predictor and outcome variables, we constructed $n \times p$ matrices with various time spans, defining the temporal resolution of time-series data and covering meaningful and regular environmental processes occurring in rivers (e.g. 24 h, 12 h, 6 h). For example, the matrices of $n \times 1$ reflected the time span of 60, 90 and 10 minutes for PR, SC and AR, respectively, whereas the matrices covering time spans of 24 hours were defined by matrices of $n \times 24$, $n \times 16$ and $n \times 96$ for PR, SC and AR, respectively. The differences in matrix $n\times$ columns of each site is a consequence of their differences in the temporal resolutions defined above. For more information about the code and functions to allow data processing and matrix construction, see `https://github.com/benoit-liquet/AD_ANN`.

Metrics for model performance and evaluation

To evaluate and compare the classification performance of AD, we calculated the four categories of the confusion matrix (i.e. true and false positives and true and false negatives; TP, FP, TN, FN, respectively) based on the discrimination threshold value t (typically 0.5; see below for detail on the threshold values we used). From these, we calculated accuracy (Acc), sensitivity (sn) and specificity (sp), and positive and negative predictive values (PPV and NPV, respectively), which allowed us to compare results directly with those of Leigh et al.[11]

$$Acc = \frac{TP + TN}{TP + FP + TN + FN}$$

$$sn = \frac{TP}{TP + FN}$$

$$sp = \frac{TN}{TN + FP}$$

$$PPV = \frac{TP}{TP + FP}$$

$$NPV = \frac{TN}{TN + FN}$$

Acc, *sn*, *sp*, NPV and PPV range between 0 and 1. *sn* assesses the probability of determining TP correctly, whereas *sp* of determining TN correctly. *Acc* assesses the ability to differentiate TP and TN correctly. Finally, PPV and NPV define the proportion of anomalous versus "normal" observations, specifically the negative and positive predictive values, respectively. When dealing with unbalanced classification problems (i.e. when there are far fewer anomalous than non-anomalous observations), many evaluation metrics are biased towards the majority class, maximising TN classification while minimising TP classification.[29,30] We therefore also calculated evaluation metrics specially formulated to provide and optimal classification for both positive and negative values in unbalanced data sets.[30] The first was balanced accuracy (*b.Acc*), which is defined as the arithmetic mean between the *sn* and *sp* values:

$$b.Acc = \frac{TP/(TP+FN) + TN/(TN+FP)}{2}$$

The second and third were the F_1 and the Matthews Correlation Coefficient (MCC), defined as:

$$f_1 = \frac{2 \times TP/(TP+FP) \times TN/(TN+FP)}{TP/(TP+FP) + TN/(TN+FP)}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (FP+TN) \times (TN+FN)}}$$

Specifically, f_1 is the harmonic mean between the sn and PPV, ranging between 1 (i.e. perfect precision and recall) and 0. MCC ranges between -1 (i.e. total disagreement between predictions and observations), 0 (i.e. random prediction) and +1 (i.e. perfect prediction and balanced ratios of the four confusion matrix categories).

Artificial Neural Networks and their implementation in time-series AD

Artificial Neural Networks (ANN) are machine learning methods, in which neurons (the basic unit of the learning process) are typically aggregated in layers that generate connections between input and output data. The typical structure of an ANN comprises input, hidden and output layers, usually connected sequentially. The input layer contains the observed data as variables (input neurons) and the output layer the predicted values (output or target neurons).[31] The input and output layers are connected by hidden layers, and the relationships between the neurons are set by non-linear activation functions. For building and checking

the performance of different types of model structure (i.e. the network structure typically associated with the number of layers and neurons), the ANN must be trained, by which the weights associated with connections between neurons are optimised using various methods and training algorithms.

A special case of ANNs adapted to time series applications are Recurrent Neural Networks (RNN), which can be considered a special case of Auto-Regressive Integrated Moving Average (ARIMA) and non-linear autoregressive moving average (NARMA) models.[32] RNN architecture mimics the cyclical connectivity of neurons, making them well suited for the analysis and prediction of non-stationary time series.[33–35] Long Short-Term Memory (LSTM) networks are a re-design of the RNNs capable of learning long-term correlations in a sequence.[36] LSTMs are structured in network units known as memory blocks composed of self-connected memory cells and three multiplicative units, namely the "input", "output" and "forget gates" connected to all cells within each memory block.[34] Unlike RNNs, LSTM avoids the "vanishing gradients problem", that is when the error signal is used to train, meaning that the network exponentially decreases the further one goes backwards in the network, by means of the gates of the network units.[37] As a result, LSTMs allow the model to be trained successfully using backpropagation through time, which is key for accounting for long-term dependent time-series sequences. Although we applied LSTM in subsequent analyses, we use the more generic term ANN from this point onwards.

In this study, ANNs were computed and fitted with "Keras", a model-level library which provides high-level building for programming for developing deep-learning models. "Keras" allows the implementation of a wide variety of neural-network building blocks (e.g. layers, activation functions, optimisers) and supports the latest and most effective advances in deep network training (including recurrent neural networks).[38] "Keras" is a high-level wrapper of "TensorFlow" and helps to provide a simplified way of building neural networks from standard types of layers while facilitating a reproducible platform for developing deep learning approaches in computational environmental sciences.[39] "Keras" is written in "Python" and

the "Keras" package provides an "R" interface to the deep-learning native functions.[40]

In each model run, we compiled a model with a pre-defined set of hyper-parameters (see below), in which the internal model parameters were iteratively updated throughout the training steps (epochs; for definitions of ANNs hyper-parameters, see Text S1 in SI). During each model run, iterations were computed until the error from the model was minimised or reached a pre-defined value. In order to assess the performance of the learning process, we used a pre-defined set of standard metrics of "Keras" commonly used for classification and regression problems. Specifically, for both training and validation data sets, we calculated the loss function for semi-supervised classification and the mean-square error and the accuracy for supervised classification. For each model run, we accounted for over-fitting using learning curves which showed how error changes as the training set size increases.[37] To do so, we compared the shape and dynamics of the learning curves of the training and validation data, which can be used to diagnose the bias and variance of the learning process. For instance, the "training learning curve" is calculated from the training data and provided information on the learning process, whereas the "validation learning curve" is calculated from the hold-out data and gives information on the model generality. For details of the dynamics of the learning curves for either semi-supervised and supervised classification with our data sets, see Supporting Information (SI). SI contains the code and R functions to allow the implementation of ANNs using "Keras" in the "R" statistical language.[41]

Optimisation of hyper-parameters and their influence on AD performance

ANNs include a broad suite of hyper-parameters that affect the ability to learn patterns from the data and the performance of model predictions. Before training, we selected ten hyper-parameters that affect (i) the network structure and (ii) the training algorithm; see SI for details on hyper-parameter definitions. Furthermore, we defined two additional hyper-parameters related to (iii) the "sliding window" defined by $n \times p$ matrices delimited at

regular temporal intervals (see above) and (iv) the "threshold classification" as a measure of the discrimination threshold value to compute the four categories of the confusion matrix (see above). For each hyper-parameter we then defined a range of values (see details in SI) that affect the performance of ANNs. The range of values depended on the type of the hyper-parameter and varied from a continuous searching space (for those hyper-parameters characterised by double precision, such as the learning rate) and discrete values (for those defined by integer values, such as number of layers or units, and functions or algorithms, such as the optimisation algorithm). In the case of the "sliding window" we decided to use discrete values, defining the temporal resolution of the recorded time-series sequence and covering meaningful and regular environmental processes occurring in rivers (e.g. 24 h, 12 h, 6 h). For more information about this issue, see SI.

Given the large number of possible models to be tested with all the combinations of hyper-parameter values, we tuned the ANN models using a Bayesian optimisation method, which is a common class of optimisation methods, especially in deep-learning networks.[42] The Bayesian optimisation method works by constructing a posterior distribution of functions (assuming a Gaussian process) associated with the variability of each hyper-parameter, which best describes the "objective function", defined as the "cost" associated with the optimisation problem. With each iteration of the algorithm, the posterior distribution improves and the algorithm becomes more accurate in those regions of the parameter space with higher likelihood to maximise or minimise the objective function.[42] The Bayesian statistical model comprises two components: (i) a Bayesian statistical model for modelling the objective function, and (ii) an acquisition function for deciding where to sample next. In our case, we applied the $mlrMBO$ toolbox implemented in the R statistical language. Compared with other black-box benchmark optimisers, the $mlrMBO$ toolbox performs well for expensive optimisation scenarios for single- and multi-objective optimisation tasks, with continuous or mixed parameter spaces.[43]

Our aim for the optimisation procedure was to find combinations of hyper-parameters

that resulted in the best performance for AD classification (see above). The optimisation procedure followed two steps, firstly generating a random search space and secondly focusing on search shrinks, based on the results generated during the random search. First, we started the algorithm by generating a random design ($n = 250$), which included a varying number of hyperparameters with the aim to generate enough variability to detect the most promising values. In our case we generated iterations by varying randomly the parameter combinations of the 12 hyper-parameters. After computing the model and the optimisation scores of each iteration, we secondly focused on search shrinks of the search space following a Bayesian optimisation method with $n = 250$ iterations based on maximising performance metrics or objective functions. In our case, we followed a multi-objective Bayesian optimisation method based on maximising, in each k iteration run, the value of $b.Acc$, f_1 and MCC scores, defined above. Such an approach allowed us to maximise the different and complementary properties summarised by each score for unbalanced classification. Following the multi-objective Bayesian method, the optimisation resulted in a total of $n = 500$ iterations (i.e. $n = 250$ for random search and $n = 250$ for Bayesian optimisation) for each combination of water-quality variables, sites and learning procedures.

To examine the effects of hyper-parameters predictor variables (hyper-parameters) on the dependent variables (performance metrics), we applied methods for causal inference using random forests.[44] Specifically, we determined the statistical importance of each hyper-parameter as a predictor on the dependent variables or optimisation scores (e.g. $b.Acc$, f_1 or MCC) by calculating the "Variable Importance" (VI). VI reflects the model performance across the entire range of predictor and response variables, converted into a set of ordinal ranks. VI can be further used to test how the model response changes as the value of any of the predictor variables is changed, meaning that VI is similar to a standard "one-parameter-at-a-time" sensitivity analysis.[45] For each hyper-parameter, we computed VI to measure the relative importance (or dependence, or contribution) of such hyper-parameter predictor variables in terms of their effect on optimisation scores. In our case, we computed

the out-of-bag error, which is an error estimation technique used to evaluate the accuracy of a random forest after permuting each predictor variable. We used the R statistical language and the "randomForest" library.[46]

For each water-quality variable and site, we retained the "best" model as that which maximised the optimisation scores. Specifically, for the complete set of candidate models, we averaged the value of $b.Acc$, f_1 or MCC and we retained the "best" model as that maximising the averaged optimisation scores. We compared the shape and dynamics of the learning curves from the "best" models (a) to diagnose whether or not the training and validation data sets were sufficiently representative (i.e. one data set could capture the statistical characteristics relative to other data sets) and (b) to test the behaviour of the learning process (i.e. underfit, overfit, good fit).

Results

Hyper-parameters optimisation for AD

The learning rate for AD stabilised early in the optimisation of ANN models, with few improvements on the performance beyond $n = 200$ model iterations (although see the ANN model for conductivity at SC). For semi-supervised classification, the costs of computing ranged from 0.312 h for turbidity in SC and 3.53 h for turbidity in AR, whereas for supervised classification the cost ranged from 1.11 h for conductivity in PR and 99.5 h turbidity in AR after $n = 500$ model iterations (see Table S2 in SI). Comparing learning methods, supervised classification had better performance and generated consistently higher values for $b.Acc$, f_1 or MCC, showing a similar and consistent pattern of the accumulative curve along the optimisation process (Fig. 2). Compared to supervised classification, semi-supervised classification required a larger number of model iterations for maximising any of the three performance metrics, notably for turbidity in SC and AR.

Considering the whole suite of model iterations, we found a consistent pattern with re-

spect to VI of those hyper-parameters affecting optimisation scores. Overall, the "Learning hyper-parameters" had higher VI values than "Model hyper-parameters" for model performance (see Tables S2 to S4 in SI). In addition, *th.class* had higher VI in semi-supervised classification, whereas *s.win* had minor VU for both supervised and semi-supervised classification. Specifically, the hyper-parameters with higher VI for *b.Acc*, f_1 and MCC, respectively, were *th.class* (78.2, 78.4 and 76.5, respectively; VI values averaged across sites, learning methods and water-quality variables), *dropout* (46.2, 47.8 and 49.0), *b.size* (43.5, 48.4 and 51.8), *momen* (31.8, 34.3 and 36.2), *l.rate* (31.3, 33.2 and 35.6). Comparing learning methods, we found that VI of hyper-parameters differed depending on the combination of types of anomalies and water-quality variables. For instance, semi-supervised classification had higher VI values for *th.class*, *dropout*, *momen* and *l.rate*, whereas supervised classification for *b.size* and *activ*. Comparing sites, *th.class* had the highest VI values in semi-supervised classification for conductivity in PR, *dropout* and *l.rate* in semi-supervised classification for turbidity in SC, *b.size* for supervised classification for turbidity in PR, *momen* for semi-supervised classification for turbidity in AR and *activ* for supervised classification for turbidity in AR. For details of the VI of those hyper-parameters independently affecting *b.Acc*, f_1 and MCC, see Tables S3 to S5 in SI.

After hyper-parameter optimisation, the "best" models also had performed well in terms of *b.Acc*, f_1 and MCC scores, for all water-quality variables and learning methods; Table 1); for details of hyper-parameters and values of the best models, see SI. However semi-supervised classification had less balanced predictions (i.e. lower values for *b.Acc*, f_1 score and MCC) than supervised classification, meaning that anomaly cases (positives) were proportionally less-correctly predicted than "normal" cases (negatives). Specifically, semi-supervised classification had lower rates of TP that we classified as true ($sn = 0.652$; values averaged across sites and water-quality variables) and lower proportions of positives and negatives that were true ($PPV = 0.512$), and that generated a moderately balanced detection rate for either TP and TN results ($b.Acc = 0.757$, $f_1 = 0.490$ and $MCC = 0.665$). By contrast, supervised

classification provided a higher and balanced detection rate for both TP and TN ($b.Acc = 0.822$, $f_1 = 0.622$ and $MCC = 0.762$), and thus anomaly cases (positives) were as predicted correctly as "normal" cases (negatives). In contrast, supervised classification had higher detection rates of TP ($sn = 0.704$) and higher rates of correct classification of TP and TN ($PPV = 0.643$). For detailed information about the performance of each model, see Table 1.

For "best" models, we additionally checked that our training and validation data sets were sufficiently representative, based on the learning process of either training or validation data. Overall, we found strong variation for each water-quality variable, site and learning method, suggesting that the presence of certain types of anomalies were not always consistent between training and validation data sets. Both for semi-supervised and supervised classification, we found that the validation data were easier to predict than training data (i.e. the "training learning curve" had poorer performance than the "validation learning curve"), suggesting that the validation data had lower complexity of anomaly types; for details of learning curves, see Fig. S2 to S5 in SI. For supervised classification, we additionally found that the validation data were unable to produce a good fit (i.e. both training and validation learning curves were almost flat), probably as a combination of the low number of anomaly cases (conductivity in PR; see Fig. 1 and Fig. S8 in SI) and the presence of a long-term anomaly event at the end of the data (turbidity in AR; see Fig. 1 and Fig. S9 in SI); the latter pattern did not happen when fitting semi-supervised classification in both data sets, for which the learning process produced a good fit for both training and validation processes. Overall, we did not detect over-fitting of the training data, a result confirmed by the relatively medium-to-low values of *dropout* (i.e. <0.5) in most of the "best" models (see Tables S2 to S4 in SI for details of the values of hyper-parameters of the "best" models).

AD among types, learning methods and sites

Anomaly types of Class 1 were present, but at low abundances, in all water-quality data sets, comprising, on average, 0.137% of cases (Table 2). By contrast, the majority of anomalies were classified as Class 3, which provided 11.6% of cases (Table 2). Anomalies of Class 3 were context-dependent with respect to each site and water-quality variable. Specifically, SC had two anomalous periods of high-variability (type E) occurring during the first four months of monitoring (see Fig. 1). PR had two drift sequences (type H) for turbidity and one period of untrustworthy data (type L) and one drift sequence (type H) for conductivity during the first period of monitoring. Finally, AR had a long-drift sequence (type H) at the end of the monitoring period.

Comparing the performance of each "best" model with respect to detecting of the different types of anomalies, we found, on average, that semi-supervised classification had higher capacity for detecting Class 1 anomalies (45.7% vs 23.6% for semi-supervised and supervised classification, respectively; averaged values across sites and water-quality variables), whereas supervised classification had proportionally higher capacity for detecting Class 3 anomalies (72.6% vs 68.8% respectively) (Table 3). However, such differences between learning methods were particularly context-dependent and based on the different combinations of types of anomalies at each site. Semi-supervised classification better detected large-sudden spikes (type A) for turbidity in AR, and small sudden spikes (J) for turbidity in PR and AR. Supervised classification, by contrast, had better performance for detecting drift (type H). Both semi-supervised and supervised classification performed well for detecting high variability (E) and untrustworthy anomalies for turbidity in SC and PR and conductivity in PR. Sudden shifts (type D) for turbidity in SC and drift (H) for conductivity in PR were not detected by any learning method.

Discussion

In this work we found that Artificial Neural Networks (ANN) provided good classification for AD in high-frequency water-quality data given their ability to deal flexibly with the challenges associated with local environmental variability. We tested the capacity of ANNs for AD under a broad range of variables, anomaly types and real-world conditions. Our data come from separate monitoring programs in different parts of the world and under contrasting environmental systems (i.e. estuarine, freshwater), so our work presents and opportunity to test the robustness and performance of ANNs for AD. We used turbidity and conductivity data, which are commonly measured by water management agencies and monitoring programs, providing an avenue to test other water quality and quantity variables in the future, which may lead to an overall increase in the performance of water-quality monitoring systems. Results of the AD showed that semi-supervised classification was able to cope better with short-term anomalies associated with a single observation or time point than supervised classification, which showed improved performance for detecting anomalies dependent on multiple context-dependent observations.

Previous work has shown that regression time-series methods are useful for AD in stationary and non-stationary time-series data sets,,[47,48] but the detection of certain anomaly types, such as sensor drift and periods of anomalously high variability, remain challenging. ANNs are a versatile method that can train models using different learning methods, and as shown by our study, can provide the flexibility required to detect a broad suite of anomaly types, including improved performance for detecting extended periods of untrustworthy data. In our case, we applied regression-based and semi-supervised ANNs similarly for AD; that is, models first predict data sequences based on "normal" cases and then classify as anomalous cases those departing from a given threshold value. Despite the good performance of ANNs for AD, our findings demonstrate that ANNs could have limitations regardless of the underlying statistical method used and their applicability in near-real time AD. In our study, c. 100 model iterations, which were necessary to obtain acceptable model perfor-

mances, required three orders of magnitude of computing time longer than regression-based ARIMA. Compared to these regression-based time-series methods, the Bayesian optimisation of hyper-parameter values allowed us to optimise classification of anomalies, at the expense of increased computing costs for hyper-parameter optimisation.

For the Australian sites, we found that the performance of semi-supervised ANNs (proposed here) and regression-based ARIMA (proposed by[11]) were equivalent in terms of correct classification, notably by providing high false-detection rates (falses) of both anomalies (positives) and "normal" cases (negatives): semi-supervised classification had higher rates of FP and FN which resulted in low values of sn and PPV. We also found that supervised ANNs considerably minimised false detection rates (i.e. low values of FP and FN and maximised sn and PPV values), when compared with the regression-based models and semi-supervised ANNs. Semi-supervised ANNs and regression-based ARIMA generated higher rates of false alarms (i.e. both FP and FN), which overestimate anomalous events, than supervised ANNs.

We did not detect over-training in the "best" models, but after checking the shape and dynamics of the learning curves we found that the training and validation data sets had inconsistent numbers and presence of anomaly types. For instance, the validation data set in turbidity in AR had a long-term anomaly event which is not present in the training data (see Fig. 1). As a result, there is inconsistency between training and validation data sets in the performance, typically the validation data had higher performance than the training data (see Figs S2 to S5 in SI). The difference in complexity was mainly a consequence of the presence of Class 3 anomalies in the data sets, usually occurring as single long-term anomalous events in each time series. Cross-validation could solve this problem, but it is also true that multiple partitioning of the data sets could have divided the time-series into multiple smaller independent time-series sequences with inconsistent representation of anomaly types among data sets. Multiple partitioning of the time-series sequence could be especially critical during the optimisation process, notably on those hyper-parameters tuning the size of the sub-samples processed during the learning process (i.e. $b.size$ and $momen$,

and also for *s.win*). All the above issues were a consequence of the proportionally lower abundance of anomalous events compared to "normal" events in the data sets, a phenomenon that makes the detection of anomalies challenging.

We found that semi-supervised ANNs were especially suited to AD of short-term anomalous events (i.e. sudden spikes, sudden shifts and small sudden spikes defined as Class 1 anomalies, following the terminology of[11]). Although semi-supervised and supervised ANNs were comparable regarding their ability to detect long-term anomalous events (i.e. drift and periods of high-variability or Class 3 anomalies), the latter had better performance in terms of correct AD (see above). At least for AD of long-term events, our results with ANNs thus outperformed those of regression-based ARIMA,[11] which is an important step forward for AD in water-quality data given such anomalies have consistently proved challenging to detect by other automated AD methods. This is because such anomalous, often very context-dependent events can behave similarly to natural, non-stationary water-quality events such that they are only detected manually by a trained eye very familiar with the local conditions. Our novel use of ANNs and Bayesian optimisation for hyper-parameter selection therefore holds much promise for AD in high-frequency water-quality data from a broad range of environments and ecosystems, including rivers, estuaries and marine waters.

Although our study is based on the analysis of data from three sites only, it provides a methodological framework for AD in high-frequency water-quality data collected from sites with contrasting non-stationary environmental processes.[49,50] We found that the non-stationary environmental conditions of each site played a substantial role in both (i) determining the types of anomalies present and (ii) increasing the uncertainty around the accurate detection of anomalies. The water-quality variables analysed here are affected by long-term environmental processes occurring at each site (i.e. seasonal precipitation patterns in both Australian and French sites), as well as regular and short-term events (e.g. cyclone floods in Australian sites and tidal regimes in the French site). Both short- and long-term environmental processes may interact and this could affected the AD performance. For Australian sites,

the detection of sudden spikes and shifts (i.e. short-term anomalies) and of drifts or periods with high variability (i.e. long-term anomalies) were rarely masked by natural environmental processes (i.e. seasonal rainfall patterns or cyclone floods), and that resulted in the optimal classification of short- and long-term anomalous events by using either semi-supervised and supervised classification, respectively (Table 3). At the French site, by contrast, strong tidal regimes occurred alongside medium-to-high seasonal discharge events, and that conditioned and limited the accuracy of detecting both short- and long-term anomalies (Table 3).

In this work, we calibrated ANNs models using Bayesian optimisation of hyper-parameter values, by means of multi-objective optimisation. Although optimisation methodology is well-established for the detection of anomalies in water-quality data,[7] multi-objective optimisation procedures have rarely been applied for AD in time-series analysis for either prediction or detection anomalies using ANNs methods; but see Perelman et al.[19] for an application of detecting anomalies using a single-objective optimisation. The application of multi-objective methodology allow us to get a balanced performance of models for detecting anomalies in a broad set of environmental conditions. Notwithstanding the potential utility of ANNs for AD demonstrated above, there is substantial room for improvement in the performance under real-world conditions.[12] First, we found that the ability for AD is context-dependent, meaning that accuracy is conditioned on the spatio-temporal environmental variability of the data set available. Anomalies are rare events, meaning that we need large data sets spanning the entire environmental variability of each locality to ensure their inclusion in a data set.[11] Second, we performed our detections based on the analysis of independent environmental variables (i.e. turbidity or conductivity), but the application of ANNs with multivariate time-series could enable us to account for the temporal correlation of multiple variables monitored at the same time.[19,24,51] Finally, methodological improvements provide additional avenues to increase the performance of ANNs in AD, such as Bayesian RNNs,[52,53] which allow quantification of the uncertainty and the use of ensemble averaging for ANNs, combining unsupervised and supervised classification[54] or the combination of

ANNs with other methodologies.[6,55]

The study of AD in high-frequency data has numerous applications in environmental and health monitoring, and fault and fraud detection, where there is a need to provide near-real time solutions and optimal performance for monitoring and to ensure the quality of data streaming. We have demonstrated that ANNs are a flexible method for providing optimal performance for AD, given they are able to cope with both long- and short-term non-stationary processes that condition high-frequency data. However, ANNs and their hyper-parameter optimisation have been understudied in the context of AD in water-quality. Given the promising results from our study, we therefore recommend further investigation and development of such methods to improve the accurate detection of a broad suite using machine learning or deep-learning methods of anomalies detection under a wide range of environmental conditions.[7] Environmental data are intrinsically variable though time and space and thus our approach is transferable for AD in complex spatio-temporal applications and ecosystems. Our findings will therefore be of relevance to water scientists and managers throughout the world in order to broaden the applicability of ANNs for efficient water monitoring.

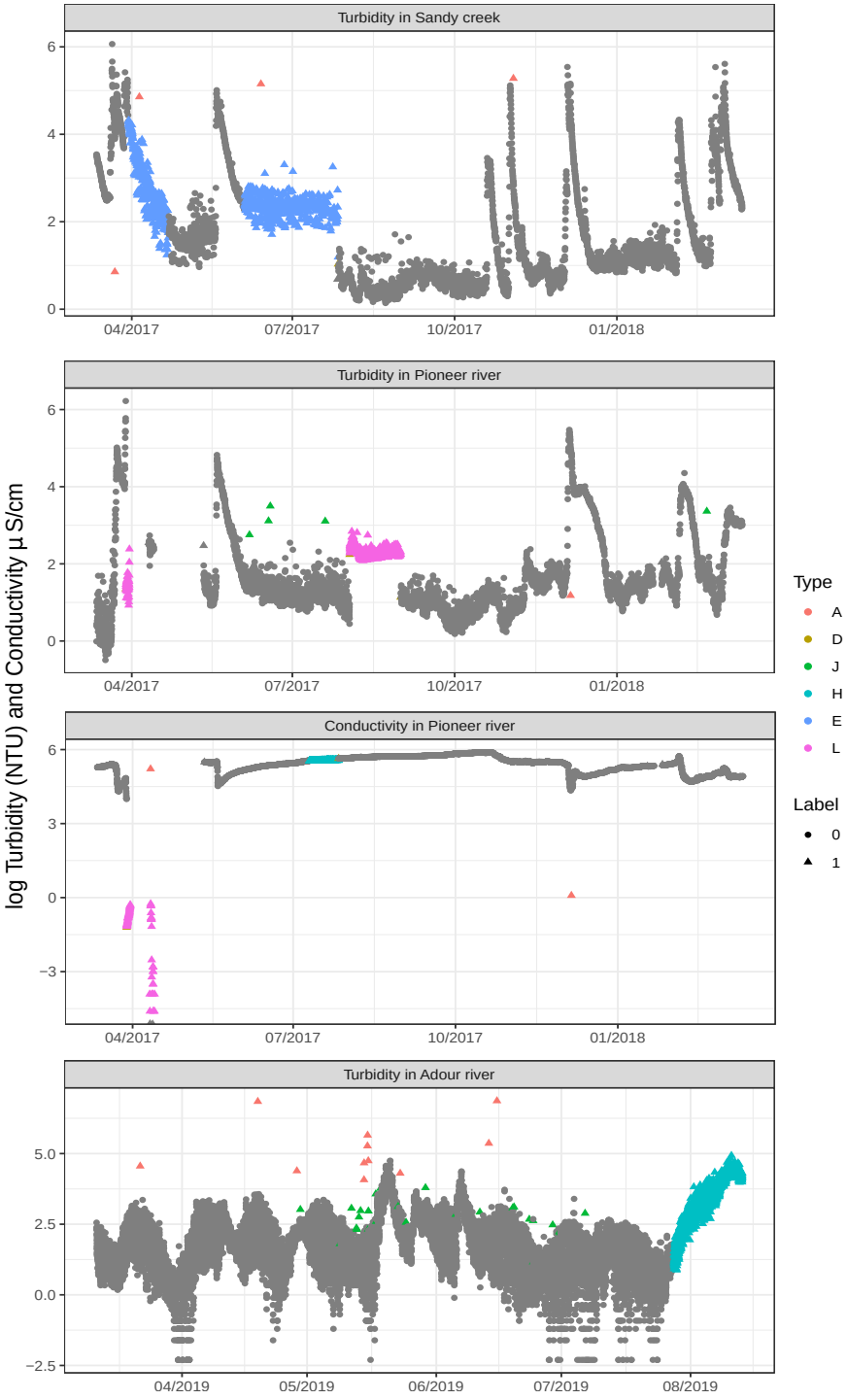

Figure 1: Observed trends for each water-quality variable and site, including the type of anomaly. Shapes correspond to the different data values (i.e. circles for "normal" values and triangles for anomalous values) and colours to the different anomaly types. Anomaly types were classified by local water-quality experts. For instance, Class 1 anomalies are defined as large sudden spikes (type A), sudden shifts (type D), small sudden spikes (type J), and Class 3 anomalies as drift (type H), high variability (type E) and untrustworthy data not defined by other types (type L).

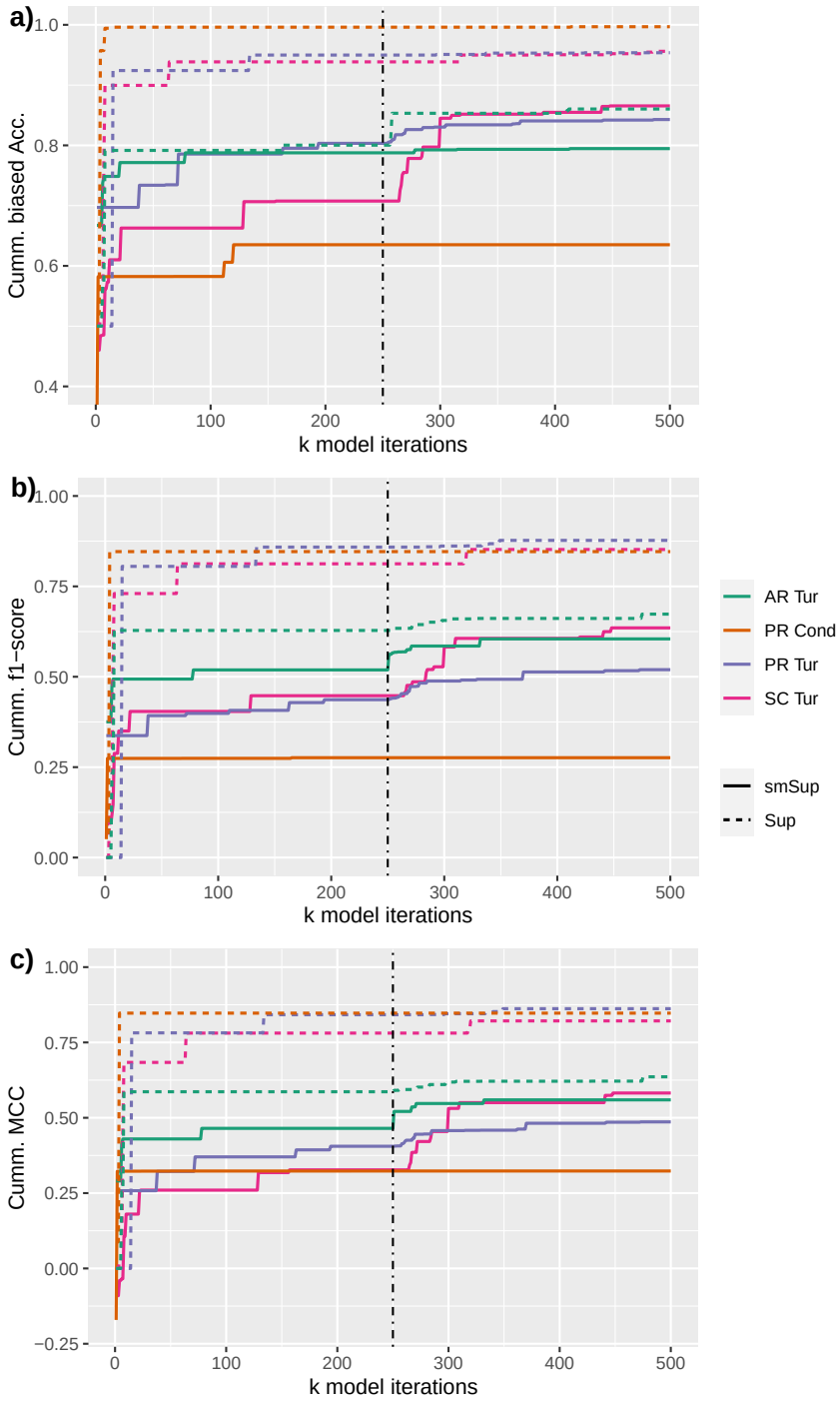

Figure 2: Cumulative optimisation scores for hyper-parameter optimisation along k model iterations. Each panel showed the procedure of multi-objective optimisation procedure of (a) balanced Accuracy, (b) f1-score and (c) Matthew's Correlation Coefficient occurring in each k iteration run. The optimisation begins searching with 250 random iterations, and then with 250 iterations of the Bayesian optimization procedure. Colours define each water-quality variable and site, whereas line patterns indicate the learning method. Abbreviations: Sandy creek (SC), Pioneer river (PR) and AR Adour river (AR); Turbidity (Tur) and Conductivity (Cond); semi-supervised classification (smSup) and supervised classification (Sup).

Acknowledgement

Funding was provided by the Energy Environment Solutions (E2S-UPPA) consortium and the BIGCEES project from E2S-UPPA ("Big model and Big data in Computational Ecology and Environmental Sciences"), the Queensland Department of Environment and Science (DES) and the ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS). A repository of the water-quality data from the in situ sensors used herein and the code used to implement methods of Artificial Neural Networks for anomaly detection are provided in the Supporting Information.

Supporting Information Available

A listing of the contents of each file supplied as Supporting Information are included. The following files are available free of charge.

- Figure S1: Scheme of the data processing for time-series AD for each learning method.
- Text S1 and Table S1: Brief definition of the hyper-parameters and range of hyper-parameter values for fitting and optimising ANN models.
- Table S2: Costs of computing time during learning process.
- Tables S3 to S5: Variable importance (VI) of values of hyper-parameters to maximise the *biased Accuracy*, the f_1-score and the *Matthews Correlation Coefficient*.
- Figures S2 to S5: Learning curves for semi-supervised and supervised learning processes for the "best" ANN models after Bayesian optimisation.
- Figures S6 to S9: Data of observed trend and the probability of anomaly detection for the "best" ANN models after Bayesian optimisation.

[2]Abbreviations: large sudden spikes (A), sudden shifts (D), small sudden spikes (J), and Class 3 anomalies defined as drift (H), high variability (E) and untrustworthy data not defined by other types (L)

Table 1: Performance of the best model for AD. For each site and water-quality variable, details of the hyper-parameters and their values of the best model are shown in SI. We calculated performance scores for ARIMA with Anomaly Detection (ArAD), semi-supervised ANNs (smSup) and supervised ANNs classification (Sup). For abbreviations of sites and variables see Figure 2.

Site	Var	Train	TN	FN	FP	TP	acc	sn	sp	PPV	NPV	b_acc	f1	MCC
SC	Tur	ArAD	4348	829	134	91	0.822	0.099	0.970	0.404	0.840	0.535	0.16	0.22
SC	Tur	smSup	1997	514	2485	405	0.445	0.441	0.446	0.140	0.795	0.443	0.21	0.40
SC	Tur	Sup	4353	540	117	374	0.878	0.409	0.974	0.762	0.890	0.692	0.53	0.64
PR	Tur	ArAD	5405	711	144	20	0.864	0.027	0.974	0.122	0.884	0.501	0.04	0.06
PR	Tur	smSup	3302	43	2180	684	0.642	0.941	0.602	0.239	0.987	0.772	0.38	0.69
PR	Tur	Sup	5330	49	208	669	0.959	0.932	0.962	0.763	0.991	0.947	0.84	0.97
PR	Cond	ArAD	5705	448	56	71	0.920	0.137	0.990	0.559	0.927	0.564	0.22	0.29
PR	Cond	smSup	2254	1	3410	480	0.445	0.998	0.398	0.123	1.000	0.698	0.22	0.60
PR	Cond	Sup	6091	0	69	48	0.989	1.000	0.989	0.410	1.000	0.994	0.58	0.65
AR	Tur	ArAD	12744	1573	452	38	0.863	0.024	0.966	0.078	0.890	0.495	0.04	0.05
AR	Tur	smSup	9540	1469	3654	142	0.654	0.088	0.723	0.037	0.867	0.406	0.05	0.07
AR	Tur	Sup	13017	925	179	495	0.924	0.349	0.986	0.734	0.934	0.668	0.47	0.55

[1]

[1] Abbreviations: True negatives (TN), false negatives (FN), false positives (FP), true positives (TP), accuracy (Acc), sensitivity (sn), specificity (sp), negative proportion of values (NPV), positive proportion of values (PPV), balanced accuracy ($b.Acc$), f_1 score (f_1) and Matthew's Correlation Coefficient (MCC).

Table 2: Number of anomalous events according to each type (columns), site and water-quality variable (rows). The anomaly types shown here were classified by local water-quality experts, and their classification is detailed in the Material and methods and in Leigh et al.[11] For abbreviations of sites, variables and training see Figure 2.

Site	Var	A	D	J	H	E	L
SC	Tur	4	1	0	0	914	0
PR	Tur	1	3	5	0	0	718
PR	Cond	2	2	0	397	0	80
AR	Tur	11	0	26	1574	0	0

[2]

Table 3: Performance of the "best" model for AD by anomaly type. For each site and water-quality variable, values represent the percentage of data values detected relative to the total number of respective anomalies labelled in the data set (see Table 3 for details). For abbreviations of sites, variables and training see Figure 2 and of anomaly types see Table 2.

	Site	Var	Train	A	D	J	H	E	L
1	SC	Tur	ArAD	1.000	0.000	-	-	0.094	-
2	SC	Tur	smSup	0.750	0.000	-	-	0.975	-
3	SC	Tur	Sup	0.750	0.000	-	-	0.986	-
4	PR	Tur	ArAD	1.000	1.000	1.000	-	-	0.010
5	PR	Tur	smSup	0.000	0.333	0.800	-	-	0.925
6	PR	Tur	Sup	0.000	0.333	0.000	-	-	0.942
7	PR	Cond	ArAD	1.000	1.000	-	0.000	-	0.362
8	PR	Cond	smSup	0.500	0.500	-	0.000	-	1.000
9	PR	Cond	Sup	0.500	0.500	-	0.000	-	0.975
10	AR	Tur	ArAD	1.000	-	0.846	0.003	-	-
11	AR	Tur	smSup	1.000	-	0.231	0.541	-	-
12	AR	Tur	Sup	0.000	-	0.038	0.728	-	-

References

(1) Horsburgh, J. S.; Jones, A. S.; Stevens, D. K.; Tarboton, D. G.; Mesner, N. O. *Environmental Modelling & Software* **2010**, *25*, 1031–1044.

(2) Rode, M.; Wade, A. J.; Cohen, M. J.; Hensley, R. T.; Bowes, M. J.; Kirchner, J. W.; Arhonditsis, G. B.; Jordan, P.; Kronvang, B.; Halliday, S. J. Sensors in the stream: the high-frequency wave of the present. 2016.

(3) Hill, D. J.; Minsker, B. S. *Environmental Modelling & Software* **2010**, *25*, 1014–1022.

(4) Horsburgh, J. S.; Reeder, S. L.; Jones, A. S.; Meline, J. *Environmental Modelling & Software* **2015**, *70*, 32–44.

(5) Jiang, J.; Wang, P.; Lung, W.-s.; Guo, L.; Li, M. *Journal of hazardous materials* **2012**, *227*, 280–291.

(6) Shi, B.; Wang, P.; Jiang, J.; Liu, R. *Science of the Total Environment* **2018**, *610*, 1390–1399.

(7) Dogo, E. M.; Nwulu, N. I.; Twala, B.; Aigbavboa, C. *Urban Water Journal* **2019**, *16*, 235–248.

(8) Chandola, V.; Banerjee, A.; Kumar, V. *ACM computing surveys (CSUR)* **2009**, *41*, 15.

(9) Gupta, M.; Gao, J.; Aggarwal, C.; Han, J. *Synthesis Lectures on Data Mining and Knowledge Discovery* **2014**, *5*, 1–129.

(10) Goldstein, M.; Uchida, S. *PloS one* **2016**, *11*, e0152173.

(11) Leigh, C.; Alsibai, O.; Hyndman, R. J.; Kandanaarachchi, S.; King, O. C.; McGree, J. M.; Neelamraju, C.; Strauss, J.; Talagala, P. D.; Turner, R. D. *Science of The Total Environment* **2019**, *664*, 885–898.

(12) Muharemi, F.; Logofătu, D.; Leon, F. *Journal of Information and Telecommunication* **2019**, *3*, 294–307.

(13) Bourgeois, W.; Romain, A.-C.; Nicolas, J.; Stuetz, R. M. *Journal of Environmental Monitoring* **2003**, *5*, 852–860.

(14) Shipmon, D. T.; Gurevitch, J. M.; Piselli, P. M.; Edwards, S. T. *arXiv preprint arXiv:1708.03665* **2017**,

(15) Makarynskyy, O.; Makarynska, D.; Kuhn, M.; Featherstone, W. *Estuarine, Coastal and Shelf Science* **2004**, *61*, 351–360.

(16) Makarynska, D.; Makarynskyy, O. *Computers & Geosciences* **2008**, *34*, 1910–1917.

(17) Wu, W.; Dandy, G. C.; Maier, H. R. *Environmental Modelling & Software* **2014**, *54*, 108–127.

(18) Tinelli, S.; Juran, I. *Water Supply* **2019**, *19*, 1785–1792.

(19) Perelman, L.; Arad, J.; Housh, M.; Ostfeld, A. *Environmental science & technology* **2012**, *46*, 8212–8219.

(20) Muharemi, F.; Logofătu, D.; Andersson, C.; Leon, F. *Modern Approaches for Intelligent Information and Database Systems*; Springer, 2018; pp 173–183.

(21) Fehst, V.; La, H. C.; Nghiem, T.-D.; Mayer, B. E.; Englert, P.; Fiebig, K.-H. Automatic vs. manual feature engineering for anomaly detection of drinking-water quality. Proceedings of the Genetic and Evolutionary Computation Conference Companion. 2018; pp 5–6.

(22) Brodie, J. *ACTFR Technical Report No. 02/03*; Australian Centre for Tropical Freshwater Research, James Cook University âĂę, 2004.

(23) Defontaine, S.; Sous, D.; Morichon, D.; Verney, R.; Monperrus, M. *Estuarine, Coastal and Shelf Science* **2019**, 106445.

(24) Leigh, C.; Kandanaarachchi, S.; McGree, J. M.; Hyndman, R. J.; Alsibai, O.; Mengersen, K.; Peterson, E. E. *PloS one* **2019**, *14*.

(25) Leigh, C.; Burford, M. A.; Connolly, R. M.; Olley, J. M.; Saeck, E.; Sheldon, F.; Smart, J. C.; Bunn, S. E. *Water* **2013**, *5*, 780–797.

(26) Wagner, R. J.; Boulger Jr, R. W.; Oblinger, C. J.; Smith, B. A. *Guidelines and standard procedures for continuous water-quality monitors: station operation, record computation, and data reporting*; 2006.

(27) Cox, B. *Science of the total environment* **2003**, *314*, 335–377.

(28) Malhotra, P.; Vig, L.; Shroff, G.; Agarwal, P. Long short term memory networks for anomaly detection in time series. Proceedings. 2015; p 89.

(29) Kelleher, J. D.; Mac Namee, B.; D'arcy, A. *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*; MIT Press, 2015.

(30) Boughorbel, S.; Jarray, F.; El-Anbari, M. *PloS one* **2017**, *12*, e0177678.

(31) Haykin, S. *Neural Networks and Learning Machines, 3/E*; Pearson Education India, 2010.

(32) Siami-Namini, S.; Tavakoli, N.; Namin, A. S. A Comparison of ARIMA and LSTM in Forecasting Time Series. 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). 2018; pp 1394–1401.

(33) Hilas, C. S.; Mastorocostas, P. A. *Knowledge-Based Systems* **2008**, *21*, 721–726.

(34) Graves, A. *Supervised sequence labelling with recurrent neural networks*; Springer, 2012; pp 37–45.

(35) Maier, H. R.; Dandy, G. C. *Environmental modelling & software* **2000**, *15*, 101–124.

(36) Hochreiter, S.; Schmidhuber, J. *Neural computation* **1997**, *9*, 1735–1780.

(37) Goodfellow, I.; Bengio, Y.; Courville, A. *Deep learning*; MIT press, 2016.

(38) Gulli, A.; Pal, S. *Deep Learning with Keras*; Packt Publishing Ltd, 2017.

(39) Rampasek, L.; Goldenberg, A. *Cell systems* **2016**, *2*, 12–14.

(40) Allaire, J.; Chollet, F. *R package version* **2017**,

(41) Team, R. C. **2017**,

(42) Snoek, J.; Larochelle, H.; Adams, R. P. Practical bayesian optimization of machine learning algorithms. Advances in neural information processing systems. 2012; pp 2951–2959.

(43) Bischl, B.; Richter, J.; Bossek, J.; Horn, D.; Thomas, J.; Lang, M. *arXiv preprint arXiv:1703.03373* **2017**,

(44) Wager, S.; Athey, S. *Journal of the American Statistical Association* **2018**, *113*, 1228–1242.

(45) Mishra, S.; Datta-Gupta, A. *Applied statistical modeling and data analytics: A practical guide for the petroleum geosciences*; Elsevier, 2017.

(46) Breiman, L. *Machine learning* **2001**, *45*, 5–32.

(47) Hyndman, R. J.; Athanasopoulos, G. *Forecasting: principles and practice*; OTexts, 2018.

(48) Box, G. E.; Jenkins, G. M.; Reinsel, G. C.; Ljung, G. M. *Time series analysis: forecasting and control*; John Wiley & Sons, 2015.

(49) Clarke, R. T. **2007**,

(50) Sivakumar, B. *Chaos in Hydrology*; Springer, 2017; pp 29–62.

(51) Sánchez-Fernández, A.; Baldán, F.; Sainz-Palmero, G.; Benítez, J.; Fuente, M. *Chemometrics and Intelligent Laboratory Systems* **2018**, *182*, 57–69.

(52) Mirikitani, D. T.; Nikolaev, N. *IEEE Transactions on Neural Networks* **2009**, *21*, 262–274.

(53) Sun, W.; Paiva, A. R.; Xu, P.; Sundaram, A.; Braatz, R. D. *arXiv preprint arXiv:1911.04386* **2019**,

(54) Comar, P. M.; Liu, L.; Saha, S.; Tan, P.-N.; Nucci, A. Combining supervised and unsupervised learning for zero-day malware detection. 2013 Proceedings IEEE INFOCOM. 2013; pp 2022–2030.

(55) Dairi, A.; Cheng, T.; Harrou, F.; Sun, Y.; Leiknes, T. *Sustainable Cities and Society* **2019**, *50*, 101670.

Graphical TOC Entry