Future Generation Computer Systems 128 (2022) 63–72

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

Handling missing data in near real-time environmental monitoring: A system and a review of selected methods

Yifan Zhang [a,b,*], Peter J. Thorburn [a]

[a] Agriculture & Food, CSIRO, Brisbane, 4067, QLD, Australia
[b] Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, Brisbane, 4072, QLD, Australia

ARTICLE INFO

Article history:
Received 4 July 2021
Received in revised form 3 September 2021
Accepted 25 September 2021
Available online 9 October 2021

Keywords:
Data imputation
Missing data
Time series

ABSTRACT

High-frequency water quality monitoring systems provide valuable measurements for predicting the trend of water quality, warning of abnormal activities or operating hydrological models. However, missing values are prevalent due to network miscommunication, device replacement or failure. Applying datasets with missing values can lead to biased results in statistical analysis or hydrological modelling work. We develop a cloud-based data processing system combining advanced algorithms to impute monitoring data in near real-time. The system provides high compatibility for supporting different water quality variables, imputation algorithms and extensive scalability to support numerous data streams. Based on the proposed approach, we review various imputation methods which can be applied to water quality data. Overall, this work provides a systematic design of a water quality data imputation system, explores the advantages and limitations of selected data imputation methods and analyses the imputation performance of two real-time water quality monitoring systems located in both the USA and Australia. The results provide practical guidelines for data imputation applications in water quality data.

1. Introduction

Real-time water quality monitoring data is valuable for constructing innovative studies that respond to dynamic temporal variations, such as water quality prediction, water quality assessment and environmental management [1]. The high-frequency measurements enable scientists to gain a comprehensive understanding of the dynamic temporal variations, which usually cannot be revealed by traditional sampling methods. However, the issue of missing data is ubiquitous in real-time water quality monitoring, which is often caused by equipment failure, network coverage or data corruption [2].

Missing data are a pervasive issue in data-driven modelling. The absence of data could cause bias in the statistical analysis, leading to invalid conclusions [3]. Moreover, the lost data makes many data modelling techniques ineffective because they presume complete information for all the variables included [4]. Hence, efficient ways of handling the missing data are urgently needed.

Common known methods to deal with missing values range from data omission to sophisticated imputation algorithms [5]. The data omission method discards samples with missing values from further analysis. Though it is easy to apply, it decreases the effective sample size dramatically. Moreover, deleting samples would cause discontinuous time-series data, which brings more difficulties in analysing temporal information.

Statistical analysis is another commonly used approach in estimating missing water quality sensor data. Kabir et al. [6] applied the mean, median and linear-based imputation methods in estimating the missing data from the water distribution network of the City of Calgary. Their experiments demonstrate that mean and median imputation tend to underestimate the variance of the data. Srebotnjak et al. [7] introduced the hot-deck imputation method to improve the European Environmental Agency water quality database. In their approach, water quality domain knowledge is critical to identify the complete samples that match the missing samples closely. Overall, for statistical-based imputation methods, integrating specific domain knowledge into the imputation process is essential to achieve promising performance.

Vast quantities of water quality sensor data afford one new opportunity for data-driven discovery. Unlike process-based models that are based on well established mathematical or physical laws, data-driven models build relationships between the

* Corresponding author at: Agriculture & Food, CSIRO, Brisbane, 4067, QLD, Australia.
E-mail addresses: yifan.zhang@uq.edu.au (Y. Zhang), peter.thorburn@csiro.au (P.J. Thorburn).

https://doi.org/10.1016/j.future.2021.09.033

system state variables without explicit knowledge of the physical behaviour of the system [8]. Ratolojanahary et al. [9] evaluated different imputation methods such as K-Nearest Neighbours (KNN), Random Forest (RF) and Multivariate Imputations by Chained Equations (MICE) for handling high rate missingness in a water quality dataset collected from France. The experimental results demonstrated that hybridization of multiple machine learning algorithms could achieve better performance than the original MICE taken alone. Kim et al. [10] compared imputation methods such as feedforward neural network and self-organizing map in estimating streamflow observations from the Taehwa River, Korea. The machine learning-based methods show promising performance in processing high-flow events. However, most data-driven methods treat water quality sensor data as a sequence of numeric values. Considering many water quality variables have predictable temporal variability, ignoring the temporal information can reduce the imputation accuracy significantly.

Neural networks with recurrent units have been widely used in processing time series data due to its capability to exhibit temporal dynamic behaviour. Many studies applied recurrent neural network such as Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) in estimating time series data. Zhang et al. [11] designed an imputation network based on GRU and residual shortcut connection. Experimental results show that the model provides higher accuracy of missing data imputation than the baseline methods. Verma and Kumar [12] proposed an accurate missing data prediction method based on the LSTM model. The LSTM based model performs well as compared to linear regression and Gaussian Process Regression on a health care dataset.

Most current studies develop imputation methods for specific water quality variables. The experimental results indicate superior performance for the proposed methods under certain circumstances. However, it is hard to determine which imputation method consistently performs best in a large spectrum of application scenarios [13]. In particular, most benchmarks do not systematically describe the workflow to process realistic monitoring data, making it hard to conduct repeatable experiments to evaluate different imputation methods.

Different from previous review studies, we firstly design a novel data imputation system that provides high compatibility in implementing various data imputation algorithms. Then, using this system, we evaluated the selected imputation algorithms on data from two real-time water quality monitoring systems. The main contributions of this article are summarized as follows.

1. We review the state-of-the-art imputation methods, analyse their advantages and limitations, and discuss the method selection in water quality data imputation. Moreover, we evaluate these selected methods on two water quality datasets collected from America and Australia.
2. We design a novel cloud-based missing data imputation system that can work with different imputation algorithms. The designed system is able to recover missing data in near real-time. It supports processing multiple water quality monitoring streams simultaneously.

The remainder of this article is organized as follows. Section 2 reviews related work and motivates the research. Section 3 presents the statistical concepts of the imputation problem. Section 4 covers a detailed description and components of the proposed imputation system. Section 5 introduced a selected number of imputation algorithms. Section 6 shows the validity as well as the penalization of the selected imputation algorithms. Sections 7 and 8 evaluate the imputation performance of the selected algorithms. Finally, Section 9 concludes the article.

2. Related work

The successive missing measurements reduce the quality and performance of real-time environmental monitoring and the efficiency of data analysis. To handle missing data in environmental monitoring, a number of data imputation methods were applied on different water quality variables.

Chen et al. [14] proposed TrAdaBoost-LSTM, which can capture the long-term dependencies among time series and leverage the related knowledge from complete datasets to fill in consecutive missing data. The results indicate that the proposed method improves the imputation accuracy by around 20% compared with alternative benchmarks. Lamrini et al. [15] applied self-organizing map (SOM)-based methods to reconstruct missing data in a drinking water treatment. Experimental results showed the efficiency and soundness of SOM algorithm. Srebotnjak et al. [7] explains the motivation and methodology of the Environmental Performance Index (EPI) Water Quality Index (WATQI) and applies hot-deck methods to impute missing WATQI in broader geographical regions. The imputation results expand the original WATQI by 39 countries to 131 countries, thereby increasing geographical coverage by 42%. Though various methods are developed to infill water quality missing data, they are evaluated only by specific water quality variables. Hence, a systematic way to implement these imputation methods to large scale monitoring data is highly needed.

Alternatively, many studies focus on comparing different imputation methods on water quality data. For example, Ratolojanahary et al. [9] combined MICE (Multivariate Imputations by Chained Equations) with Random Forest (RF), Boosted Regression Trees (BRT), K-Nearest Neighbours (KNN) and Support Vector Regression (SVR) to address the issue of data imputation, in the context of water quality assessment. The results showed that MICE-SVR is the best in that it converges faster than the three others and provides the best performance. Betrie et al. [16] compared three imputation methods such as iterative robust model-based imputation (IRMI), multiple imputations of incomplete multivariate data (AMELIA), and sequential imputation for missing values (IMPSEQ) on infilling water quality data collected from copper–molybdenum–gold–silver–rhenium mine site. The results showed that IMPSEQ and IRMI are suitable to impute missing values in water-quality databases at mine sites, whereas AMELIA is not. Tabari and Talaee [17] examined the efficiency of the multilayer perceptron (MLP) and radial basis function (RBF) networks for recovering the missing values of 13 water quality parameters based on data from five stations located along the Maroon River, Iran. It was also found that the MLP models were superior to the RBF models to reconstruct water quality missing data. To the best of our knowledge, most of the benchmark studies do not cover the advanced neural network-based imputation models. Considering deep neural network models outperforms traditional imputation methods in many studies [2,18–21], ignoring this type of method cannot provide the comprehensive performance evaluation for imputing water quality data under realistic use cases.

In contrast to the previous studies, in this paper we propose a cloud-based missing data imputation system that can support different imputation algorithms. This offers the user a systemic approach to run imputation tasks on large scale water quality monitoring data. Furthermore, we review a select number of imputation methods, which covers the statistical-based, data-driven model-based and neural network-based solutions.

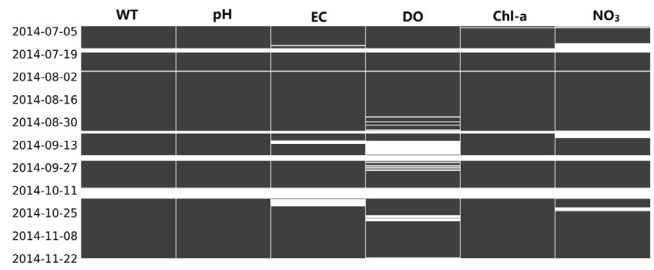


Fig. 1. An example of missing data in the water quality monitoring system. This system measures seven water quality variables such as water temperature (WT), pH, electric conductivity (EC), dissolved oxygen (DO), chlorophyll-a (Chl-a) and nitrate (NO_3). The while gap and grey block represent the missing data and available data.

3. Water quality data missing problem

The quality of statistical analytic can be highly affected by the proportion of missing data [22]. According to the study proposed by Rubin [23], missing data are often categorized into the following three types:

- Missing Completely At Random (MCAR). If every measurement in the dataset has the same probability of being missing, the datasets is defined to be missing completely at random. This implies that the causes of the missing data are unrelated to the data. MCAR is an ideal assumption but it rarely occurs in practice.
- Missing At Random (MAR). Suppose only groups of measurements in the datasets have the same probability of being missing, and the observed data define the probability. In that case, we define the dataset to be missing at random. MAR is a more general and realistic assumption than MCAR. Under this assumption, the missingness can be modelled by using the observed data.
- Missing Not At Random (MNAR). This refers to the case when neither MCAR nor MAR holds. When the dataset is MNAR, the fact that the data are missing is systematically related to the unobserved data. It is hard to handle this missing data type because it will require strong assumptions about the patterns of missingness.

It is worth noting that missing water quality data tends to follow the MAR mechanism [24]. Hence, it is reasonable to estimate missing water quality information by applying different analytical and modelling approaches.

Fig. 1 illustrates how data was missing in East Russell River station from North Queensland, Australia [25]. This station is part of the Great Barrier Reef catchment loads monitoring program, which will be described in Section 7.1. As can be seen, each variable had monitoring data missed. Among these variables, some variables such as DO, NO_3 and EC have a larger number of missing data than other variables. Missing a consecutive number of data cross multiple variables bring significant challenges in estimating water quality monitoring data accurately.

4. Data imputation system

In this paper, we designed and implemented a novel missing data imputation system. Unlike previous studies designed for specific imputation algorithms [26,27], we abstracted the critical processing steps in data imputation and built the system based on modular design principles (Fig. 2).

The system uses PyTorch as the backend engine for deep learning models and python imputation package like impyute [28] as the backend engine for none deep learning algorithms. The system is developed on the Amazon Cloud to scale up for hundreds of environmental monitoring data streams.

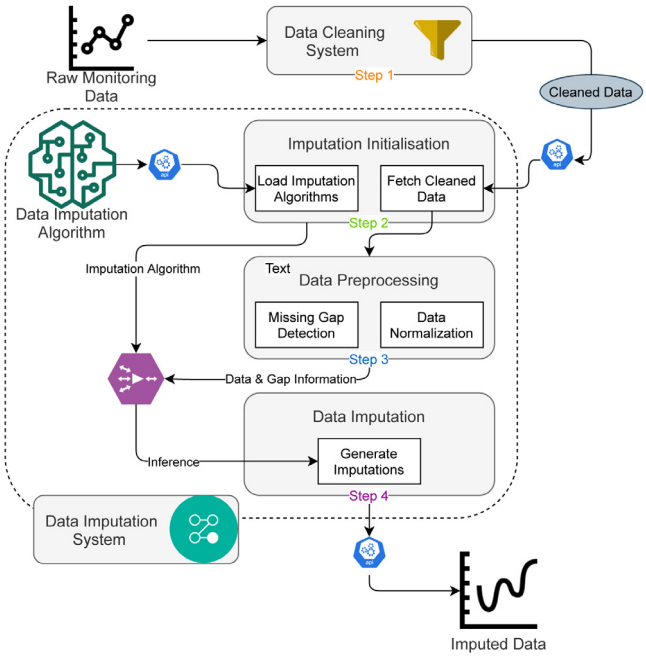


Fig. 2. Cloud-based Data Imputation System. The system takes raw water quality measurements as inputs. It works with a data cleaning system to remove obvious outliers. After selecting the data imputation algorithm and fetching the cleaned data, a data processing step is required to generate inputs for the algorithm. Followed by this, the model infills missing gaps accordingly.

A detailed explanation of the main steps illustrated in Fig. 2 for using the system is:

1. Water Quality Monitoring Data Cleaning: The inputs of the system are raw monitoring data steams. There are many noises, or error information got involved during the data collection. Considering data imputation algorithms need good quality inputs, we first pass the input datasets through a data cleaning system. It provides basic cleaning functions by using thresholds filtering and sensor reference value check. The water quality experts set maximum and minimum thresholds for different variables to exclude outliers. Many sensor manufacturers such as NICO [29] provide the sensor reference value to indicate the quality of the measurement. Hence, it is also used to remove invalid measurements. In this system, every data stream is attached with its metadata configuration, and the data cleaning process is automatically triggered when there are enough incoming data collected.
2. Imputation System Initialization: In this step, the system loads the selected data imputation algorithm. Each data stream can apply different imputation methods regarding its configuration. Also, the cleaned data are fetched through the data API during the initialization.
3. Data Preprocessing: Data normalization is essential for data-driven modelling. It rescaled different input variables into the same range, which is necessary for measurements with different units. Moreover, the system also labelled all the gaps in the input data. Followed by this, a sliding window strategy is applied to generate the algorithms' inputs.
4. Imputation Data Generation: With the selected data imputation algorithm and input data, the imputation value

can be generated for each missing gaps. With the gap information collected in step 3, the system can infill value to the corresponding gaps and output the final complete data.

In the design above, the data cleaning system ensures that there are no outliers in the raw data. It is essential for data imputation because outliers with extreme values can heavily mislead the imputation results [30]. In the data imputation system, each imputation method works as a plugin, providing flexibility to extend the system in supporting more advanced imputation algorithms.

Based on this proposed system, we can implement different data imputation algorithms and evaluate their performance according to the imputed results.

5. Overview of imputation methods

In this section, we listed and described popularly used imputation methods in estimating water quality data. They can be classified into three groups: statistical-based methods, model-based methods and neural network-based methods.

5.1. Statistical based

Deterministic imputation replaces the missing data with plausible values, which can be derived by substituting values from the available observed variables [31]. Here, we listed three popular used imputation methods: mean imputation, last observation carried forward (LOCF) and linear imputation.

5.1.1. Mean imputation

Mean imputation involves replacing the missing value with the arithmetic mean of all the other available values.

$$\hat{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

5.1.2. Last observation carried forward

Last Observation Carried Forward (LOCF) is commonly used for dealing with missing values. In this method, the missing value is imputed from the last observation in the dataset. This method makes the unrealistic assumption that there is no change at all since the last measured observation [32]. LOCF method is often used in dealing continuous value under MCAR condition.

5.1.3. Linear imputation

Linear interpolation estimates missing values based on the adjacent available values. It is preferred for estimating continuously missing data over a short time interval. For a missing value y_i, linear interpolation generates the estimation based on the closest preceding and succeeding available values y_h and y_j,

$$\hat{y_i} = y_h + \frac{y_j - y_h}{x_j - x_h}(x - x_h), x_h < x_i < x_j. \tag{2}$$

where x_h, x_i and x_j represent the preceding, current and succeeding value of x.

Linear imputation is simple, fast, and requires only two available samples to impute each missing data period. On the other hand, the accuracy of Linear imputation typically decreases as the length of the missing data period increases.

5.2. Model based

Model-based imputation aims to build the predictive models for each target variable that contains missing values. Several commonly-used imputation methods are explained in this subsection. This includes Expectation-Maximization, Multiple imputations by chained equations, and the k-nearest neighbour.

5.2.1. Expectation-maximization

Expectation-Maximization (EM) is a parametric method to impute missing values based on the maximum likelihood estimation. The EM generates estimated values for missing data through the expectation and maximization steps.

In the expectation step, the missing data is estimated based on all the observed data and the current estimate model parameters. Mathematically, the calculation can be expressed as:

$$Q(\theta|\theta^i) = \int l(\theta|Y) f(Y_{mis}|Y_{obs}, \theta^i) dY_{mis} \tag{3}$$

where $l(\theta|Y)$ is log likelihood function of complete data, $l(\theta|Y_{obs})$ is log likelihood function of observed data, and $f(Y_{mis}|Y_{obs}, \theta)$ is the predictive distribution of missing data given θ.

In the maximization step, the expectation of the complete data log likelihood from the previous estimation step is maximized to help get the next guess of θ:

$$\theta^{i+1} = \underset{\theta}{argmax} Q(\theta|\theta^i) \tag{4}$$

This two-step process iterates until convergent and the missing data can be finally estimated.

5.2.2. Multiple imputations by chained equations

Multiple imputations by chained equations (MICE) is one principled method of addressing missing data. The three stages of MICE are described as below:

In the beginning, all missing values are filled by random sampling data with replacement from the existing values. The first variable with missing values x_1 is generated based on all other variables $x_2, \ldots, x_k$, restricted to individuals with the observed x_1. Missing values in x_1 are replaced by simulated draws from the corresponding posterior predictive distribution of x_1. Then, the next variable with missing values x_2 is regressed based on all other variables $x_1, x_3, \ldots, x_k$, restricted to individuals with the observed x_2, and using the imputed values of x_1. Again, missing values in x_2 are replaced by draws from the posterior predictive distribution of x_2. The process is applied to all other variables with missing values. This procedure would repeated for several turns to produce a single imputed dataset.

5.2.3. K-nearest neighbours

The k-nearest neighbour (KNN) is a popular approach in data processing applications. It is designed to replace missing values by using k-most similar non-missing data. A categorical missing value is imputed with the majority among its k nearest neighbours, and the numerical missing value is filled by calculating the average value of the k nearest neighbours.

To select k number of nearest neighbours, the similarity between the data and its nearest neighbours should be maximal. Various distance functions have been used to measure the distance between data A and B. The Euclidean distance function is selected in most studies. For example A = $(x_1, x_2, \ldots, x_m)$ and B = $(y_1, y_2, \ldots, y_m)$, where m is the feature space dimensionality. To calculate the distance between points A and B, the normalized Euclidean metric is calculated as:

$$\text{dist}(x_i, x_j) = \sqrt{\frac{\sum_{p=1}^{n}(x_i^p - x_j^p)^2}{n}} \tag{5}$$

In addition, the usually used method is Minkowski distance (or its variants) as follows:

$$\text{dist}(x_i, x_j) = \left(\sum_{p=1}^{n} |x_i^p - x_j^p|^r\right)^{\frac{1}{r}} \tag{6}$$

where r is a non-negative integer called Minkowski coefficient. Minkowski distance is regarded as Manhattan distance while $r = 1$ and as Euclidean distance while $r = 2$.

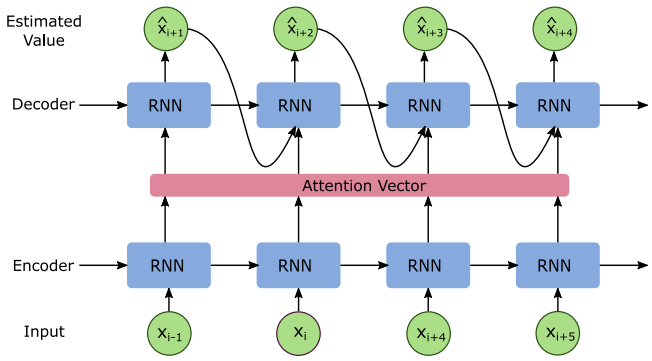


Fig. 3. Sequence-to-Sequence model for data imputation.

5.3. Neural network based

Deep Neural Networks (DNNs) have recently achieved state-of-the-art performance on various speech recognition and computer vision tasks. This motivates one to investigate applying DNNs in estimating missing data. Here, we listed data imputation models with two mainstream architectures: Sequence-to-Sequence Model, and Recurrent Neural Network.

5.3.1. Sequence-to-sequence model

Sequence-to-sequence learning emerges as an effective paradigm for dealing with variable-length inputs and outputs. It aims to directly model the conditional probability $p(y|x)$ of mapping an input sequence, $x_1, \ldots, x_n$, into an output sequence, $y_1, \ldots, y_m$ [33]. This process is undertaken by the encoder–decoder framework proposed by Cho et al. [34]. Since missing values can randomly happen during the data collection, generating an arbitrary number of estimated values is important.

Fig. 3 illustrates the standard sequence-to-sequence model for imputation tasks. In Fig. 3, the encoder computes a representation s for each input sequence. Based on this input representation, the decoder generates an output sequence, one unit at a time. In this approach, the conditional probability is decomposed as:

$$\log p(y|x) = \sum_{j=1}^{m} logp(y_j|y_{<j}, x, s) \tag{7}$$

where x and y are the input and output sequences, s denotes the representation for each input sequence.

When it comes to the imputation problem, the input is the sequence of available data points around the missing gap. The output of the model is the estimation of values at each time index of the missing gap. In the sequence-to-sequence model, recurrent neural network (RNN) units are applied in the encoder and decoder components to process time series data. Also, the attention mechanism enables the model to learn how to focus on a specific range of the input sequence for the differing outputs.

1. SSIM

   Sequence-to-Sequence Imputation Model (SSIM) [2] is the first data imputation model based on the sequence-to-sequence architecture and attention mechanism. SSIM uses the long short-term memory network (LSTM) to capture the available temporal information between gaps, and the global attention mechanism is applied to let SSIM focus on specific parts of the input for estimating different missing values.

2. Dual-SSIM

   The Dual-SSIM [18] extends the conventional SSIM model by having two separate encoders to process the input sequence around the missing gap. The Gated Recurrent Unit (GRU) based encoders can split the information before and after the missing gap naturally. Hence, no extra gap vectors need to be created to locate the missing values' locations. In addition, the global attention mechanism is extended to support processing temporal representations learned from two different encoders.

5.3.2. Recurrent neural network

Instead of imputing missing values directly, another idea is to estimate missing data when computing other correlated prediction tasks. This idea is popularly applied in practical applications such as health care and biology [19]. By combining the imputation process with prediction tasks, the accuracy of estimating missing values can be improved with the guide of the correlated prediction tasks.

In this approach, a recurrent component and a regression component work together for generating imputation outputs. Usually, the recurrent component is achieved by a recurrent neural network and the regression component is achieved by a fully-connected network. A standard recurrent network can be represented as:

$$\mathrm{h}_t = \sigma(W_h \mathrm{h}_{t-1} + U_h \mathrm{x}_t + b_h) \tag{8}$$

where σ is the sigmoid function, W_h, U_h and b_h are model parameters, and h_t is the hidden state of previous time steps.

For the imputation problem, x_t may have missing values so one cannot use x_t as the input directly as in the above equation. Instead, many studies [19–21] use a 'complement' input x_t^c when x_t is missing. x_t^c can be calculated in different strategies such as mean of the dataset, same as its last measurement or others or intermediate outputs inside the model. Here, we take the third strategy as an example. In this case, the input x_t^c can be calculate as follows:

$$\hat{x}_t = W_x \mathrm{h}_{t-1} + b_x, \tag{9}$$

$$x_t^c = \mathrm{m}_t \odot x_t + (1 - \mathrm{m}_t) \odot \hat{x}_t \tag{10}$$

where $\hat{x}_t$ is the estimated vector based on the hidden state h_{t-1}. $\odot$ represents the element-wise multiplication. m_t denotes whether the input is missing at time step t. W_x, b_x are model parameters.

Fig. 4 depicts the customized recurrent neural network model for handling time series with missing values. In this approach, the inputs with missing values are used to train the model for the prediction task. To get the prediction results, the model needs to estimate the missing values as the intermediate outputs.

In order to help the model identify the gaps in the inputs, additional time interval and masking vectors need to be provided as supplementary information. Time interval vector is designed to measure the distance for each variable since its last observation. Masking vector m is applied to indicate which variables are missing at time step t.

1. BRITS

   BRITS [20] is a recurrent neural network based method for missing value imputation in time series data. By adapting the recurrent neural network, BRITS treats missing values as variables of the computation graph and updates the estimations during the backpropagation process. In this way, gradients information from both forward and backward directions are used together to update the missing value, which leads to a more accurate estimation.

   BRITS uses a data-driven imputation procedure to estimate missing data. To use BRITS, the data imputation task

Table 1
Strength and limitations of the methods listed in Section 5.

Method	Strength	Limitation
Mean Imputation		The variability in the data is reduced;the standard deviations and the variance estimates may get underestimated [35],
LOCF	Easy to understand, Efficient to apply	The assumption is mostly unrealistic [36],
Linear Imputation		There should be a linear relationship between the predictor and response variables
EM	Good Interpretability,	High risk to overfit the training data
MICE	Lazy Learning, do not build	Computationally expensive for large datasets
KNN	models from training data	
SSIM	Can capture and use the	Black box method,
Dual-SSIM	temporal information,	Performance heavily relies on
BRITS	Deep architecture brings	hyperparameters tuning,
M-RNN	strong representation learning	High computational cost to build the model

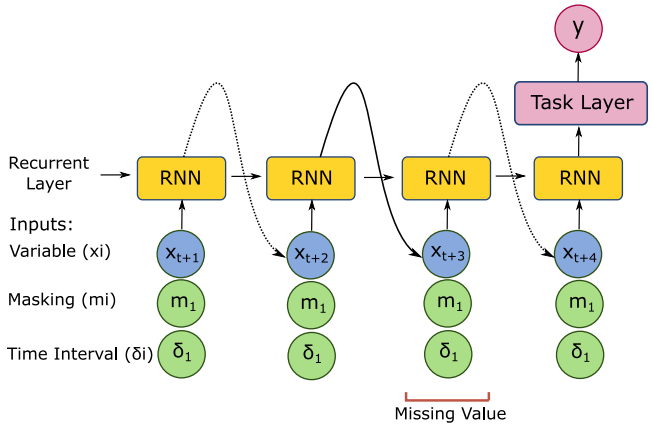


Fig. 4. Customized recurrent neural network model for data imputation.

and prediction tasks are usually implemented jointly. Experiments demonstrate that this strategy can boosts the imputation accuracy and classification task's accuracy significantly.

2. M-RNN

   M-RNN [21] is a Multi-directional Recurrent Neural Network. It uses information within the same data stream to interpolate data, and also imputes missing values across data streams. It contains an interpolation block and an imputation block. The interpolation block constructs an interpolation function that operates within the data stream. The imputation block constructs an imputation function that operates across streams. M-RNN overperforms several benchmarks on five real-world medical datasets.

6. Advantages and limitations

In this section, we summarized the main relative advantages and limitations of these imputation techniques, as well as their suitability to modelling purpose (see Table 1).

1. Statistical-based Methods

   When using statistical-based methods, missing values are replaced by a value defined by a certain rule. This approach is computationally simple but ignores the relationship between variables in the datasets. Hence, it often underestimate the variability because each unobserved value carries the same weight in the analysis as the known observed values [37]. Moreover, some statistical-based methods assume all the missing data follow a constant pattern. For example, they are all close to the medium value (Mean Imputation) or the preceding available value (LOCF). Therefore, the statistical-based methods are often potentially biased and should be used with great caution [38].

2. Model-based Methods

   Model-based imputation takes into account the relationship between different variables by building regression models for missing features that take the non-missing features as inputs [39]. However, it has significant computational overhead. First, the run-time cost of applying the model on large scale datasets can be prohibitive. Second, it depends heavily on the type and nature of the data, and cannot be used as an out-of-the-box pre-processing step.

3. Neural Network-based Methods

   Neural networks with recurrent architecture have the ability to capture long-term temporal dependencies and variable-length observations, which cannot be supported by other imputation modelling technologies. However, neural
   network-based models have the following limitations: (1) due to the multilayer nonlinear structure, neural network models are often criticized for being non-transparent and the outputs not traceable by humans [40]. (2) with the rapid increase in the volumes of data, the time required to train the neural network is increasing accordingly [41]. (3) compared with traditional machine learning algorithms, deep learning is highly dependent on hyper-parameter tuning [42]. Hence, getting insights into the neural network's mechanism is necessary to provide trustworthy imputation solutions.

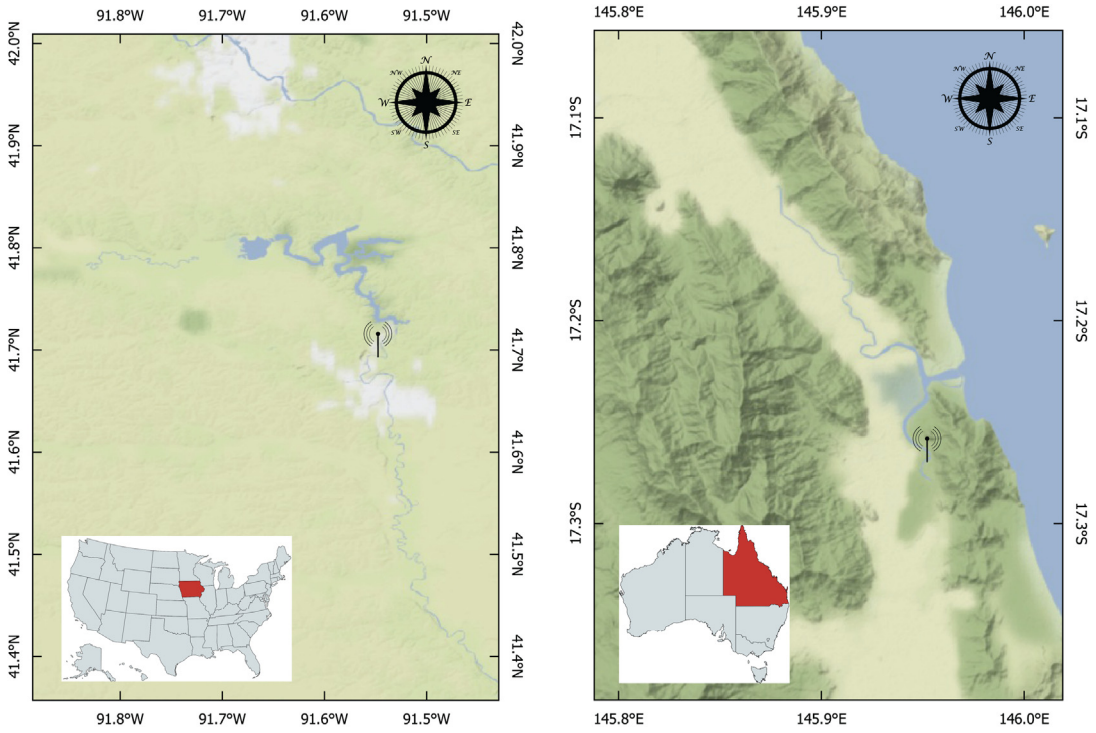
7. Experimental cases

7.1. Water quality monitoring networks

In the experimental section, we choose to evaluate the data imputation methods by using water quality measurements collected from two water quality monitoring systems located in both the USA and Australia (see Fig. 5).

7.1.1. Iowa water quality information system

The Iowa Water Quality Information System (IWQIS) [43] is a water quality monitoring network across the state of Iowa, USA. It offers real-time measurements for water quality variables such as NO_x concentration, pH, turbidity, electric conductance, dissolved oxygen, and temperature. The data used in this section were collected from one monitoring station located in Clear Creek Watershed.



(a) The Iowa Water Quality Information System (IWQIS). The black icon presents the selected monitoring station located in the upstream of Iowa River.

(b) The Great Barrier Reef catchment loads monitoring program. The black icon represents the in-situ monitoring station located in the upstream of Russell River.

Fig. 5. Two water quality monitoring systems in the USA and Australia.

Table 2
Water quality data collected from two monitoring systems.

Variables	Unit	Min	Max	Mean	Std Dev
IWQIS					
Water Temperature	°C	2.2	29.6	18.9	7.4
Conductivity	μS/cm	265.8	683.7	554.7	68.9
Nitrate	mg/L	1.9	14.9	7.8	2.5
GBR					
Water Temperature	°C	15.7	33.7	24.4	3.4
Conductivity	μS/cm	61.9	1021.4	403.3	205.9
Nitrate	mg/L	0.002	3.3	0.3	0.3

7.1.2. GBR catchment loads monitoring program

The Great Barrier Reef catchment loads monitoring program aims to help track long-term and short-term water quality trends in North Queensland, Australia [44]. The program monitors all intensive land use catchments. It includes 43 monitored sites across 20 key catchment areas for monitoring sediments and nutrients, and 20 sites for pesticides.

7.2. Water quality monitoring data

Three water quality variables such as water temperature, conductivity and nitrate are measured in both systems (Table 2). In the following section, we choose to estimate the missing data for water temperature and nitrate concentration. The sensor data was normalized and cleaned to remove obvious outliers. In addition, we resampled the water quality measurements to one hour time interval in this experiment. In addition, all three variables are used as inputs when imputing water temperature and nitrate concentration. One thing worth mentioning here is that we design to impute missing measurements in a few hours because the inputs data are collected hourly. When weekly, monthly or yearly data are fed, the models in this paper could generate imputations at corresponding time scale, respectively.

To evaluate the imputation accuracy, we first prepare the training and testing data without missing values by selecting different periods from the data mentioned in Table 2.

Then, a sliding window algorithm is used to generate all the samples. For each sample, we mask a consecutive number of data as the groundtruth so we can measure how good the imputation results are.

7.3. Evaluation metrics

We evaluate the performance of recovering missing data based on the root mean square error (RMSE) and mean absolute error (MAE).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(|f_i - \hat{f}_i|)^2}, \tag{11}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - \hat{f}_i|, \tag{12}$$

where f_i and $\hat{f}_i$ are the true and estimated values of a water quality variable under monitoring, respectively.

8. Results

8.1. Water temperature data

Table 3 compares the performance of estimating the missing water temperature for ten different imputation methods. Both RMSE and MAE are used in quantifying the imputation accuracy. It is clear that the Dual-SSIM achieves the best performance for both RMSE and MAE in two different datasets. For example, the

Table 3
Imputation Accuracy For Water Temperature with Gap Size 6.

	Water Temperature			
	USA		AU	
	RMSE	MAE	RMSE	MAE
Mean Imputation	0.48 (± 0.041)	0.4 (± 0.063)	0.409 (± 0.062)	0.348 (± 0.05)
LOCF	0.479 (± 0.042)	0.399 (± 0.064)	0.413 (± 0.06)	0.351 (± 0.048)
Linear Imputation	0.48 (± 0.04)	0.4 (± 0.062)	0.413 (± 0.062)	0.353 (± 0.051)
EM	0.48 (± 0.042)	0.4 (± 0.064)	0.417 (± 0.052)	0.356 (± 0.04)
MICE	0.48 (± 0.041)	0.4 (± 0.063)	0.409 (± 0.062)	0.348 (± 0.05)
KNN	0.649 (± 0.015)	0.583 (± 0.038)	0.552 (± 0.166)	0.483 (± 0.227)
Dual-SSIM	**0.004 (± 0.001)**	**0.004 (± 0.001)**	**0.015 (± 0.004)**	**0.013 (± 0.004)**
SSIM	0.007 (± 0.004)	0.007 (± 0.004)	0.031 (± 0.017)	0.028 (± 0.016)
BRITS	0.007 (± 0.004)	0.006 (± 0.003)	0.03 (± 0.009)	0.022 (± 0.008)
M-RNN	0.026 (± 0.002)	0.021 (± 0.001)	0.066 (± 0.007)	0.056 (± 0.004)

Dual-SSIM gets 0.004 and 0.015 RMSE scores in processing the data collected from the USA and AU monitoring systems, respectively. The next best imputation models are neural network-based methods like SSIM, BRITS and M-RNN which outperform both the statistical and model-based solutions significantly. The results in Table 3 demonstrate that neural network-based imputation methods are able to utilize the strong temporal patterns appeared in many water quality variables.

On the contrary, imputation methods such as mean imputation, LOCF and linear imputation did not perform well in imputing missing water temperature data. These methods ignore the temporal patterns, and most of them assume a linear relationship happened among the water quality variables. Hence, they have low imputation accuracy in the experiments.

8.2. Nitrate data

Compared to water temperature, the nitrate concentration does not follow a clear changing pattern in the daily or weekly time scale. The temporal variations of the nitrate concentration can only be identified when checking the trend throughout several months. Hence, it is very challenging in estimating missing nitrate measurements.

Table 4 summarizes the imputation accuracy for the nitrate concentration measured in both the USA and AU monitoring systems. In this experiment, the Dual-SSIM still performs the best for both RMSE and MAE scores. It has 0.002 and 0.041 RMSE scores for both USA and AU data, respectively. The imputation accuracy varies among these two datasets. The two monitoring systems are running in different climate zones and affected by various types of land use and agricultural activities. Hence, the nitrate concentration would not always follow a similar pattern, and the imputation model should be specific for different water quality datasets.

When imputing nitrate data, some straightforward methods such as mean imputation can get better performance than modelling methods like linear imputation, EM and KNN. For example, mean imputation has an RMSE score of 0.007 when applied to the USA data, while linear imputation and EM get 0.012 and 0.023 RMSE scores, respectively. This result indicates that data-driven modelling methods may not achieve the expected performance if the temporal information is not fully utilized.

In Fig. 6, we compared the imputation results of all the methods analysed in this paper. The concentration of nitrate fluctuated during this period, which is usually driven by the fertilizing activities. In this example, 6 successive data points around the peak are missing. The gap is infilled by using different imputation methods. It is evident that LOCF, Mean and KNN all generate a straight line, indicating poor imputation accuracy. Linear imputation is also not suitable when values vary during the period. Among the neural network-based methods, the Dual-SSIM generates imputations with the correct trend. Most methods underestimate the range of missing values significantly.

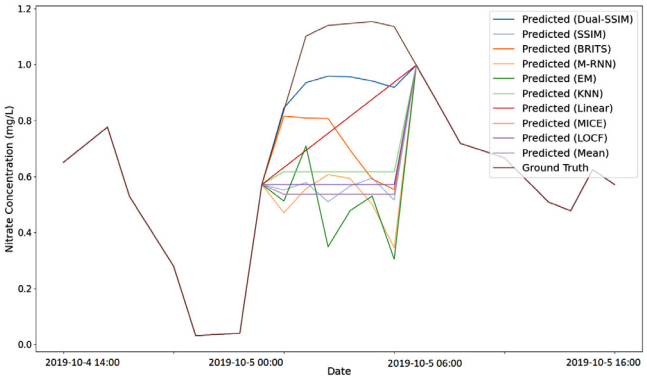


Fig. 6. Model outputs in imputing 6 consecutive missing values for nitrate concentration from GBR monitoring network. The solid red-brown line represents the ground truth data. Other lines represent the imputation results generated by different models. 20 available data before and after the gap are used as the model's input.

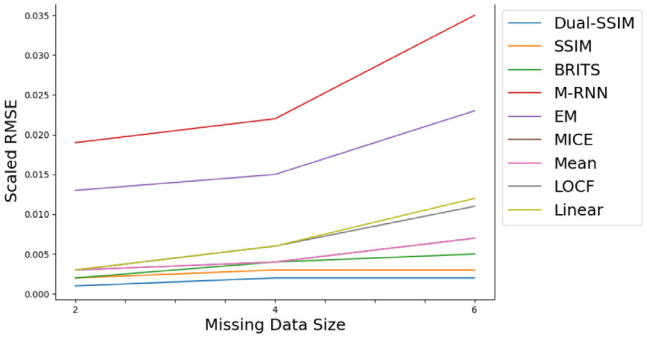


Fig. 7. Imputation accuracy for estimating data with different size by different methods.

8.3. Effect of data gap size

The size of missing water quality data may affect the imputation accuracy significantly. The longer the data gaps are, the less help the available data can provide. Hence, we evaluated how the imputation accuracy changes when estimating missing data with different size.

Fig. 7 shows how the imputation model performs when estimating missing data with different size. In the experimental setting, nine imputation methods were evaluated on missing gaps of size 2, 4 and 6, respectively. In general, the imputation bias increases when the size of the missing data is growing.

The imputation methods have close performance when dealing with missing data of size 2. Supported by a large amount of

Table 4
Imputation Accuracy For Nitrate Concentration with Gap Size 6.

	Nitrate			
	USA		AU	
	RMSE	MAE	RMSE	MAE
Mean Imputation	0.007 (± 0.002)	0.005 (± 0.001)	0.149 (± 0.07)	0.096 (± 0.05)
LOCF	0.011 (± 0.003)	0.007 (± 0.002)	0.155 (± 0.07)	0.084 (± 0.04)
Linear Imputation	0.012 (± 0.001)	0.007 (± 0.001)	0.144 (± 0.044)	0.086 (± 0.027)
EM	0.023 (± 0.006)	0.015 (± 0.003)	0.194 (± 0.089)	0.122 (± 0.058)
MICE	0.007 (± 0.002)	0.005 (± 0.001)	0.149 (± 0.07)	0.096 (± 0.05)
KNN	0.712 (± 0.123)	0.661 (± 0.166)	0.552 (± 0.166)	0.483 (± 0.227)
Dual-SSIM	**0.002 (± 0.001)**	**0.002 (± 0.001)**	**0.078 (± 0.067)**	**0.069 (± 0.06)**
SSIM	0.003 (± 0.001)	0.003 (± 0.001)	0.098 (± 0.083)	0.089 (± 0.076)
BRITS	0.005 (± 0.002)	0.004 (± 0.002)	0.117 (± 0.107)	0.077 (± 0.072)
M-RNN	0.035 (± 0.021)	0.029 (± 0.017)	0.269 (± 0.009)	0.233 (± 0.044)

available information, most imputation methods can handle the small amount of missing data efficiently.

When doubling the gap size to 4, the neural network-based methods still offer high accuracy for imputing missing data. In this case, the neural network-based models can still learn the temporal patterns from the available data, which provide a strong guide to infill the missing data. On the contrary, linear imputation, LOCF and EM show dramatical performance downgrade.

By keep increasing the size of missing data, the imputation accuracy of neural network-based methods still maintains on the low level. However, statistical and model-based methods could not achieve encouraging performance. For time-series data, the missing value has less relevance to the available information at time steps far away from the gap. If the imputation model does not have a strong capability in capturing the temporal patterns from the available data, imputing missing data without ground-truth at nearby time steps can be difficult.

9. Conclusion

This paper provides a review of popularly used data imputation methods and qualitatively compares their respective advantages and disadvantages of being used in water quality measurements.

The imputation techniques listed in this paper can be grouped into three different types. The statistical-based imputation methods infill missing data based on statistical analyses of the time series data. The model-based methods infill missing value using regression models. The neural network-based methods build specific neural network models in predicting the missing data. Imputation methods built on different mechanisms have their own advantages and limitations. Hence, suitable imputation methods need to be chosen for specific circumstances.

In summary, the size of the missing data affects the imputation accuracy significantly. Most imputation methods perform well in infilling missing data in a short period. For instance, only 1 or 2 data are missing in one period. When there are a greater number of missing values in the datasets, imputation methods which cannot utilize the temporal information have significantly poorer performance. Because they benefit from the recurrent architecture, the neural network-based methods show the promising results in processing datasets with large gaps.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

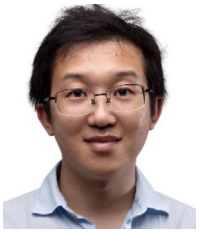
Acknowledgement

We would like to thank the Great Barrier Reef catchment loads monitoring program and Iowa water quality information system for providing valuable real-time water quality monitoring datasets. This work was conducted within the CSIRO Digiscape Future Science Platform.

References

[1] Y. Zhang, P. Fitch, P.J. Thorburn, Predicting the trend of dissolved oxygen based on the kPCA-RNN model, Water 12 (2) (2020) 585.
[2] Y. Zhang, P. Thorburn, W. Xiang, P. Fitch, SSIM -A deep learning approach for recovering missing time series sensor data, IEEE Internet Things J. 6 (4) (2019) 6618–6628.
[3] H. Kang, The prevention and handling of the missing data, Korean J. Anesthesiol. 64 (5) (2013) 402.
[4] M. Soley-Bori, Dealing with missing data: Key assumptions and methods for applied analysis, Boston Univ. 4 (2013) 1–19.
[5] C. Nieh, S. Dorevitch, L.C. Liu, R.M. Jones, Evaluation of imputation methods for microbial surface water quality studies, Environ. Sci. Process. Impacts 16 (5) (2014) 1145–1153.
[6] G. Kabir, S. Tesfamariam, J. Hemsing, R. Sadiq, Handling incomplete and missing data in water network database using imputation methods, Sustain. Resilient Infrastruct. (2019) 1–13.
[7] T. Srebotnjak, G. Carr, A. de Sherbinin, C. Rickwood, A global water quality index and hot-deck imputation of missing data, Ecol. Indic. 17 (2012) 108–119.
[8] D.P. Solomatine, A. Ostfeld, Data-driven modelling: some past experiences and new approaches, J. Hydroinform. 10 (1) (2008) 3–22.
[9] R. Ratolojanahary, R.H. Ngouna, K. Medjaher, J. Junca-Bourié, F. Dauriac, M. Sebilo, Model selection to improve multiple imputation for handling high rate missingness in a water quality dataset, Expert Syst. Appl. 131 (2019) 299–307.
[10] M. Kim, S. Baek, M. Ligaray, J. Pyo, M. Park, K.H. Cho, Comparative studies of different imputation methods for recovering streamflow observation, Water 7 (12) (2015) 6847–6860.
[11] J. Zhang, X. Mu, J. Fang, Y. Yang, Time series imputation via integration of revealed information based on the residual shortcut connection, IEEE Access 7 (2019) 102397–102405.
[12] H. Verma, S. Kumar, An accurate missing data prediction method using LSTM based deep learning for health care, in: Proceedings of the 20th International Conference on Distributed Computing and Networking, ACM, 2019, pp. 371–376.
[13] S. Jäger, A. Allhorn, F. Bießmann, A benchmark for data imputation methods, Front. Big Data 4 (2021).
[14] Z. Chen, H. Xu, P. Jiang, S. Yu, G. Lin, I. Bychkov, A. Hmelnov, G. Ruzhnikov, N. Zhu, Z. Liu, A transfer learning-based LSTM strategy for imputing large-scale consecutive missing data and its application in a water quality prediction system, J. Hydrol. (2021) 126573.
[15] B. Lamrini, E.-K. Lakhal, M.-V. Le Lann, L. Wehenkel, Data validation and missing data reconstruction using self-organizing map for water treatment, Neural Comput. Appl. 20 (4) (2011) 575–588.
[16] G.D. Betrie, R. Sadiq, S. Tesfamariam, K.A. Morin, On the issue of incomplete and missing water-quality data in mine site databases: Comparing three imputation methods, Mine Water Environ. 35 (1) (2016) 3–9.
[17] H. Tabari, P. Hosseinzadeh Talaee, Reconstruction of river water quality missing data using artificial neural networks, Water Qual. Res. J. Canada 50 (4) (2015) 326–335.

[18] Y. Zhang, P.J. Thorburn, A dual-head attention model for time series data imputation, Comput. Electron. Agric. 189 (2021) 106377, http://dx.doi.org/10.1016/j.compag.2021.106377.
[19] Z. Che, S. Purushotham, K. Cho, D. Sontag, Y. Liu, Recurrent neural networks for multivariate time series with missing values, Sci. Rep. 8 (1) (2018) 6085.
[20] W. Cao, D. Wang, J. Li, H. Zhou, L. Li, Y. Li, BRITS: bidirectional recurrent imputation for time series, in: Advances in Neural Information Processing Systems, 2018, pp. 6775–6785.
[21] J. Yoon, W.R. Zame, M. van der Schaar, Estimating missing data in temporal data streams using multi-directional recurrent neural networks, IEEE Trans. Biomed. Eng. (2018).
[22] Y. Dong, C.-Y.J. Peng, Principled missing data methods for researchers, SpringerPlus 2 (1) (2013) 222.
[23] D.B. Rubin, Inference and missing data, Biometrika 63 (3) (1976) 581–592.
[24] C. Güler, G.D. Thyne, J.E. McCray, K.A. Turner, Evaluation of graphical and multivariate statistical methods for classification of water chemistry data, Hydrogeol. J. 10 (4) (2002) 455–474.
[25] QLD, Great barrier reef real time water quality data, 2020, https://www.kaggle.com/ivivan/real-time-water-quality-data. (Accessed 20 July 2020).
[26] J. Lin, N. Li, M.A. Alam, Y. Ma, Data-driven missing data imputation in cluster monitoring system based on deep neural network, Appl. Intell. 50 (3) (2020) 860–877.
[27] L. Shen, P.R. Stopher, A process for trip purpose imputation from Global Positioning System data, Transp. Res. C 36 (2013) 261–267.
[28] E. Law, 2017, https://impyute.readthedocs.io/. (Accessed 4 August 2021).
[29] NICO, 2018, https://www.hydrotechzs.com/downloads/nico-manual.pdf. (Accessed 4 August 2021).
[30] N. Kumar, M.A. Hoque, M. Shahjaman, S. Islam, M.N. Mollah, A new approach of outlier-robust missing value imputation for metabolomics data analysis, Curr. Bioinform. 14 (1) (2019) 43–52.
[31] J. Nissen, R. Donatello, B. Van Dusen, Missing data and bias in physics education research: A case for using multiple imputation, Phys. Rev. Phys. Educ. Res. 15 (2) (2019) 020106.
[32] P.R. Houck, S. Mazumdar, T. Koru-Sengul, G. Tang, B.H. Mulsant, B.G. Pollock, C.F. Reynolds III, Estimating treatment effects from longitudinal clinical trial data with missing values: comparative analyses using different methods, Psychiatry Res. 129 (2) (2004) 209–215.
[33] M.-T. Luong, Q.V. Le, I. Sutskever, O. Vinyals, L. Kaiser, Multi-task sequence to sequence learning, 2015, arXiv preprint arXiv:1511.06114.
[34] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, 2014, arXiv preprint arXiv:1406.1078.
[35] J.D. Dziura, L.A. Post, Q. Zhao, Z. Fu, P. Peduzzi, Strategies for dealing with missing data in clinical trials: from design to analysis, Yale J. Biol. Med. 86 (3) (2013) 343.
[36] A.M. Wood, I.R. White, S.G. Thompson, Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals, Clin. Trials 1 (4) (2004) 368–376.
[37] J.C. Jakobsen, C. Gluud, J. Wetterslev, P. Winkel, When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowcharts, BMC Med. Res. Methodol. 17 (1) (2017) 162.
[38] A.W. Jørgensen, L.H. Lundstrøm, J. Wetterslev, A. Astrup, P.C. Gøtzsche, Comparison of results from different imputation techniques for missing data from an anti-obesity drug trial, PLoS One 9 (11) (2014).
[39] I. Žliobaitė, J. Hollmén, Optimizing regression models for data streams with missing values, Mach. Learn. 99 (1) (2015) 47–73.
[40] V. Buhrmester, D. Münch, M. Arens, Analysis of explainers of black box deep neural networks for computer vision: A survey, 2019, arXiv preprint arXiv:1911.12116.
[41] D. Justus, J. Brennan, S. Bonner, A.S. McGough, Predicting the computational cost of deep learning models, in: 2018 IEEE International Conference on Big Data, Big Data, IEEE, 2018, pp. 3873–3882.
[42] X. Zhang, X. Chen, L. Yao, C. Ge, M. Dong, Deep neural network hyper-parameter optimization with orthogonal array tuning, in: International Conference on Neural Information Processing, Springer, 2019, pp. 287–295.
[43] IOWA, 2019, http://iwqis.iowawis.org/. (Accessed 20 August 2019).
[44] QLD, Reef 2050 water quality improvement plan, 2018, https://www.reefplan.qld.gov.au/. (Accessed 20 July 2018).



Yifan Zhang (S'08-M'13) received the B.Eng. and M.Eng. degrees in systems engineering from Beijing University of Aeronautics and Astronautics, Beijing, China in 2008 and 2011, respectively, and the Ph.D. degree in data science from Queensland University of Technology, Brisbane, Australia, in 2016. From 2017 to 2020, he worked as a Postdoctoral Fellow with CSIRO, Agriculture & Food. His work focused on applying machine learning in processing spatiotemporal environmental sensor data. The applications include water quality forecasting, sensor data imputation and spatiotemporal feature learning. He is currently a Postdoctoral Fellow in the University of Queensland, QAAFI. His research interests include artificial intelligence, time series modelling, image processing, cloud computing and the internet of things.

Peter J. Thorburn is an agricultural scientist with strong multidisciplinary interests in the dynamics of soil–plant interactions and a strong commitment to enhancing the sustainability of agricultural systems. With a background in soil science and plant physiology, his work focuses on developing and applying simulation models to understand soil and plant interactions in agricultural production systems, aiming to determine management systems that can reduce detrimental environmental impacts while, still continuing to produce significant economic and social outcomes in current and future climates. His research has won international awards and has had substantial policy impact. He is a Senior Principal Research Scientist within CSIRO's Agriculture and Food Business Unit. He is the CSIRO representative on the APSIM Imitative, the joint venture that owns the APSIM farming systems model. He also coordinates research on agriculture's effect on the Great Barrier Reef across CSIRO's Agriculture and Food Business Unit, and co-leads the strategic collaboration between CSIRO Agriculture and Food and AgResearch (New Zealand). His international activities include co-leading, with Professor Boote from the University of Florida, the Crop Modelling stream of the Agricultural Modelling Improvement and Intercomparison Program (AgMIP), an international collaboration with American and European institutions developing increased capacity for quantitatively assessing climate change impacts on global food security. He is also on the Science Advisory Group of New Zealand's OVERSEER agricultural nutrient management model.