Machine Learning Approaches to Improve and Predict Water Quality Data

Yi-Fan Zhang [a], **Peter J. Thorburn** [a], **Maria P. Vilas** [a] **and Peter Fitch** [b]

[a]*CSIRO, Agriculture & Food, Brisbane, QLD, 4067, Australia*
[b]*CSIRO, Land & Water, Canberra, ACT, 2601, Australia*
Email: yi-fan.zhang@csiro.au

Abstract: Changes in water quality have a variety of economic impacts on human and ecosystem health. The widespread use of in-situ high-frequency monitoring instrumentation enables a better characterisation of water quality processes, leading to more meaningful decision making. The large amount of data collected by the high-frequency sensors creates new opportunities for machine learning methods to better understand data-intensive processes in aquatic ecosystems and improve data streams coming from sensors.

CSIRO's DigiscapeGBR project aims to help protect the Great Barrier Reef (GBR) by enabling upstream sugarcane growers to make better nitrogen fertiliser management decisions, supporting water quality improvements which are critical to meet ecological targets for protecting the health of the Reef's ecosystems. Current studies based on machine learning in the DigicapeGBR project to improve data and predict future water quality are mainly focused on the following:

- Water quality prediction.

  The development of reliable water quality predictions is critical for improving the management of aquatic ecosystems. Predicting the response of coupled biogeochemical and physical systems is challenging due to the complexity and non-linearity of these systems. Thus, a machine learning approach may be accurate in predicting water quality as it accounts for non-linearity.

- Water quality data imputation

  Missing data are unavoidable in water quality monitoring systems. Most data analysis methods require complete data as inputs. Incomplete data can produce biased or wrong results, with negative effects on the conclusions drawn from the water quality data. Classical methods for filling gaps in the data perform poorly when consecutive data points are missing. Thus, there is a need to compare the performance of a machine learning approach against classical imputation methods.

- Water quality outlier detection

  The data collected by environmental sensors can be noisy and have outliers due to sensor malfunction. These anomalies make the data more difficult to analyse and interpret. Therefore, the identification of atypical observations is an essential concern in water quality monitoring. Typical methods for outlier detection have low detection rates given the high variability in water quality data. Thus, there is a need to investigate the performance of a machine learning approach for outlier detection in water quality data.

In this paper, we introduce and summarise the machine learning based modelling work we have been investigating for solving the three challenges described above. For water quality prediction, neural network models based on artificial neural network (ANN), recurrent neural network (RNN) and convolutional neural network (CNN) have been developed to forecast changes in dissolved oxygen (DO) and other water quality variables in rivers draining to the GBR. For water quality data imputation, we proposed a sequence-to-sequence imputation model (SSIM) for recovering missing data in high-frequency monitoring systems. The SSIM uses the state-of-the-art sequence-to-sequence architecture, and the Long Short Term Memory Network (LSTM) is chosen to utilise both the past and future information for a given time. For water quality outlier detection, we have been investigating neural network models combined with the wavelet decomposition. All the models show promising results in solving some of the challenges around water quality data management and prediction.

Keywords: *Artificial intelligence, neural network, data driven, time series, great barrier reef, nitrate*

1 INTRODUCTION

Deteriorating water quality has a variety of social, environmental and economic impacts. The widespread use of in-situ water quality monitoring sensors has provided researchers with a wealth of water quality data, making data-driven modelling a possibility. Compared with process-based modelling methods, machine learning approaches do not require prior physical, chemical or biological knowledge of the system being monitored. Instead, they try to ”learn” the hidden patterns in the data by processing the huge amount of monitoring data directly. Being able to reproduce these patterns not only allows prediction of water quality but also can be used to improve data streams through overcoming issues such as erroneous data.

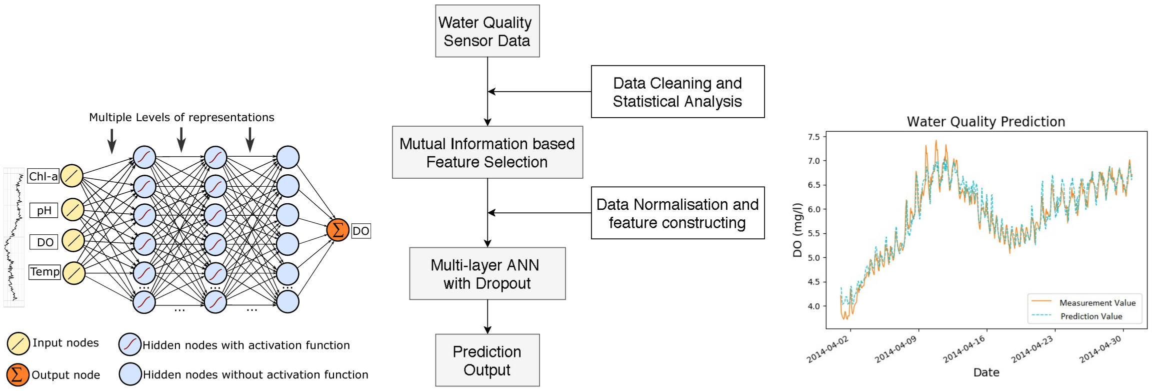
This paper explores the application of machine learning approaches in aquatic sciences. This specific context of this work is to better understand and improve the quality of water discharged from coastal catchments into GBR ecosystems (Thorburn et al. (2019)). This work was undertaken within the CSIRO Digiscape Future Science Platform GBR Project (CSIRO (2019)).

2 APPLICATIONS OF MACHINE LEARNING IN WATER QUALITY

2.1 PREDICTION

Predicting the trend of water quality is a significant challenge in several fields of study such as prawn aquaculture (Dabrowski et al. (2018)) and pollution management (Chang et al. (2015)). Predicting water quality can guide the implementation of management actions to maintain good water quality conditions (Thorburn and Wilkinson (2013)).

Data-driven models have proved successful in predicting water quality. For instance, artificial neural networks (ANN) have been used to successfully predict the chemical oxygen demand during river restoration in Wuxi city, China (Ruben et al. (2018)). Given that changes in water quality are driven by the physical environment and complex biogeochemical processes (Vilas et al. (2017)), data-driven predictions need to include variables that account for both the physical and biogeochemical environment.


Multiple Levels of representations
Chl-a
pH
DO
Temp
DO
Input nodes
Output node
Hidden nodes with activation function
Hidden nodes without activation function
Water Quality Sensor Data
Data Cleaning and Statistical Analysis
Mutual Information based Feature Selection
Data Normalisation and feature constructing
Multi-layer ANN with Dropout
Prediction Output
Water Quality Prediction
DO (mg/l)
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
2014-04-02
2014-04-09
2014-04-16
2014-04-23
2014-04-30
Date
Measurement Value
Prediction Value

(a) Multi-layer artificial neural network for predicting DO.

(b) The workflow of predicting DO by using ANN and MI.

(c) 90 mins ahead DO prediction in the wet season (1/4/2014-30/4/2014).

Figure 1. Water quality prediction model based on multi-layer ANN and MI.

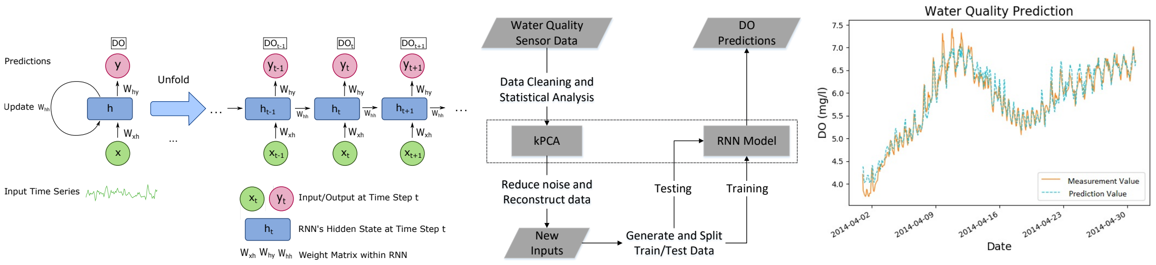
ANN based model. We proposed a predictive water quality model based on multi-layer artificial neural network and mutual information (MI) (Zhang et al. (2019)). MI is used to evaluate and choose the most relevant water quality input variables by taking into account the non-linear relationships between the variables. A multi-layer ANN model is built to learn the levels of representations and approximate complex regression functions (Figure 1a).

Compared to other ANN-based modelling work (Sarkar and Pandey (2015)), we proposed a systematic way to select appropriate water quality inputs for the specific water quality predictive task (Figure 1b). Unlike the Pearson correlation coefficient which is commonly used to determine the relationship between variables, MI is more general and contains information about all linear and non-linear dependencies between variables. In addition, the multi-layer ANN with the dropout mechanism is proved to have superior capabilities to traditional shallow neural networks in preventing overfitting and capturing non-linear temporal correlations (Hinton et al. (2012)).

In this study, water quality data collected from Baffle Creek, Australia was used in an experiment to test the accuracy of our ANN model. Climate in North Queensland is characterized by wet and dry seasonal patterns. In general, the wet season spans from November to April and the dry season spans from May to October. In the wet season test illustrated in Figure 1c, the concentration of dissolved oxygen (DO) increases from around 4.0 mg L^{-1} to nearly 7.5 mg L^{-1} within the first month and thereafter fluctuates between 5.5 and 7.0 mg L^{-1}. During this period, our multi-layer ANN model is effective in predicting the diurnal pattern as well as the long-term variability. Also, the multi-layer ANN model has a quick response when DO concentrations start to change.

Our model had superior R^2 scores for predicting DO 90 mins or 120 mins into the future from the last observed data in both the dry and wet season, compared to single layer ANN, support vector regressor and linear regression models. The results indicate that our multi-layer ANN model can provide accurate predictions for the trend of DO in the upcoming hours and is a useful supportive tool for water quality management in aquatic ecosystems.

RNN based model. Recurrent Neural Networks (RNN) are able to exhibit a dynamic temporal behaviour by establishing connections between units form a directed cycle. Compared to a feed-forward neural network, RNN has information travelling in both directions. Computations derived from the earlier input are fed back into the network, which is critical in learning the non-linear relationships between multiple water quality parameters.



(a) Recurrent neural network for predicting DO.

(b) The workflow of predicting DO by using the kPCA-RNN model.

(c) 3 hours ahead predicting for the concentration of DO.

Figure 2. Predicting the trend of dissolved oxygen (DO) based on kPCA-RNN model.

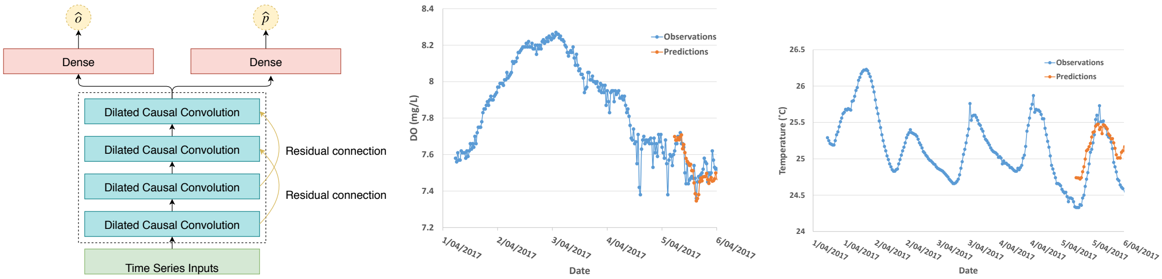
We proposed a predictive water quality model based on a combination of a kernel principal component analysis (kPCA) and recurrent neural network (RNN) (Figure 2a). Water quality parameters are reconstructed based on kPCA method, which aims to reduce the noise from the raw sensory data and preserve redundant information. With the RNN's recurrent connections, our model can dynamically operate on input information as a trace of acquired previous information.

As exhibited in Figure 2a, the structure of the RNN model across time can be expressed as a deep neural network with one layer per time step. Because this feedback loop occurs at every time step in the series, each hidden state contains traces not only of the previously hidden state but also of all past hidden states as long as memory can persist. The workflow is depicted in Figure 2b. Firstly, the kPCA method is implemented on the collected water quality data. Principal components are constructed and used as new inputs to the RNN model. After training and testing the RNN model, the concentration of DO in the upcoming time steps can be estimated.

We evaluated our kPCA-RNN model on DO data collected from Burnett River, Australia. The kPCA-RNN model achieved R^2 scores up to 0.91, 0.82 and 0.67 for predicting the concentration of DO in the upcoming 1, 2 and 3 hours, respectively. In the 3 hours ahead prediction (Figure 2c), around 93 % of the results were within $\pm 10\,\%$ range of the original observations.

CNN based model. As well as recurrent based neural networks, we also investigated using convolutional neural network (CNN) based modelling techniques for long-term water quality prediction.

Compared to RNN, Temporal Convolutional Networks (TCN) can capture longer-range patterns using a hierarchy of temporal convolutional filters (Lea et al. (2017)). Instead of using recurrent structure to maintain temporal dependencies, the TCN applies various size of convolutional filters to obtain the temporal dependencies at different time scales. Also, the dilated convolutions (Van Den Oord et al. (2016)) increase the receptive field significantly so time series data with long historical observations can be fully used.



(a) The multi-task temporal CNN for predicting water quality.

(b) Predicting the trend of DO in the following 1 day.

(c) Predicting the trend of Temperature in the following 1 day.

Figure 3. Multi-task temporal convolutional network for predicting water quality.

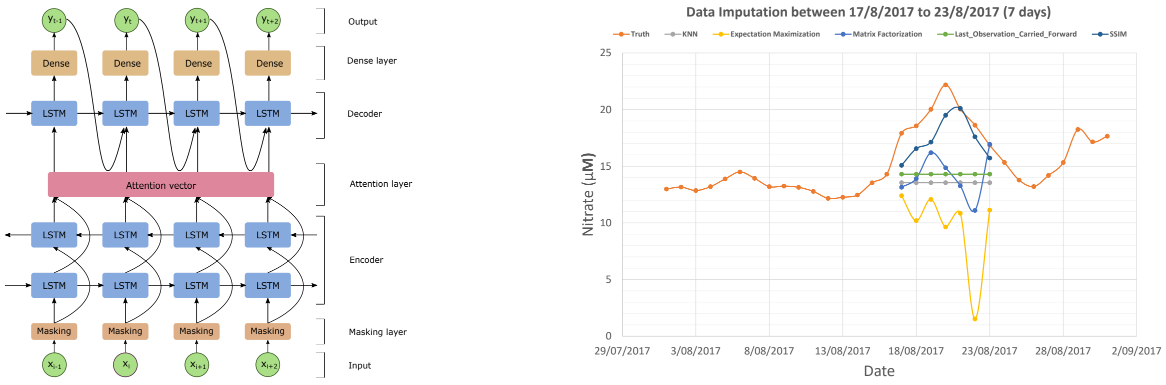
We proposed a multi-task temporal convolution network (MTCN) for predicting multiple water quality variables (Zhang et al. (2019)). The MTCN is able to forecast various water quality constituents simultaneously (Figure 3a). This enables knowledge sharing between multiple learning processes, and also reduces the required computing resources significantly. By adjusting the dilation factors and filter size, the MTCN can cover a wide range of time series data by applying a hierarchy of filters with various size. In addition, the residual connections help to maintain the stability of the deep neural network by enhancing the information flow through the initial layers to last layers in the deep neural network. The task-specific dense layers with a linear activation function are added on top of the shared convolutional layers. Each dense layer is designed to focus on learning the task-specific knowledge and generate the estimations for each of the water quality variables separately.

Water quality data from the Burnett River was chosen to test the MTCN. In the experiment, the MTCN is able to simultaneously forecast changing temperature and DO in the following two days (Figures 3c and 3b). Instead of predicting various water quality variables independently, this multi-task learning approach forces the model to extract the correlation between various water quality variables explicitly, which can better make use of the prior knowledge of the system. As a result, the MTCN yields half hourly predictions in the following day, so 48 outputs are generated concurrently. Compared to predictive models that produce a single or a few number of predictions, the MTCN allows to capture diurnal changes in water quality. This gives managers more time to put in place management operations.

2.2 IMPUTATION

Missing data are unavoidable in long-term and real-time monitoring networks due to issues such as network communication outage, sensor failure or lack of maintenance. Although multiple methods have been proposed for filling gaps in the data, most methods give poor estimates when multiple data points are missing. The greater number of missing data points, the more difficult the gap to fill. To address this issue, some methods reconstruct missing data based on other variables collected at the same time. When all variables have gaps in the data, these methods cannot be applied. The performance of deep learning methods have shown promise for data imputation. However, these methods rely highly on a large volume of training data. In many scenarios, it is difficult to obtain large volumes of data from monitoring networks.

Hence, we proposed a new sequence-to-sequence imputation model (SSIM) for recovering missing data in sensor networks (Zhang et al. (2019)). The SSIM uses the state-of-the-art sequence-to-sequence deep learning architecture. In conjunction with Long Short Term Memory Network (LSTM), the memory and attention



(a) SSIM architecture with the attention mechanism.

(b) Example data imputation from the SSIM method and four traditional imputation methods.

Figure 4. Sequence-to-sequence Imputation Model (SSIM).

mechanisms utilize both the past and future information for a given time. In addition, a variable-length sliding window algorithm is developed to generate a large number of training samples from small data sets so that the SSIM can be trained with small data sets.

The SSIM utilizes the sequence-to-sequence architecture with the attention mechanism as depicted in Figure 4a, where the encoder and decoder are two key functional components. The encoder processes an input time series and maps it to a high-dimensional vector. The decoder takes input from the vector and yields target data sequences. Also, the attention mechanism enables the decoder to learn how to focus on a specific range of the input sequence for the differing outputs.

An example of the application of the SSIM is shown in Figure 4b. The missing data points are predicted by the SSIM one by one from 17/8/2017 to 23/8/2017. Each time the model yields one output, it will combine this output with the previous inputs to generate the next new output. The SSIM utilises the available information both from the past and future time steps, which enhances model's ability to capture the trend through a period. Processing information from two directions can efficiently reduce accumulated predictive error.

Experimental results were presented to demonstrate that the proposed model can recover missing data sequences more accurately than other benchmark methods, such as ARIMA, Seasonal ARIMA, Matrix Factorization, Multivariate Imputation by Chained Equations and Expectation Maximization. The SSIM is therefore a promising approach for filling gaps in the data obtained from wireless sensor networks. The SSIM has been implemented into a cloud-based data imputation system for processing water quality sensor data in real-time.

2.3 OUTLIER DETECTION

The third problem addressed in our work is outlier detection. Identification of atypical observations is an important element of water quality monitoring (Di Blasi et al. (2013)). The data collected by environmental sensors can be noisy and have outliers due errors in the sensors or physical interference. These anomalies make the data more difficult to analyse and interpret and have a significant impact on the implementation of water quality management actions.

Though various outlier detection algorithms have been proposed, most of them cannot achieve good accuracy when dealing with high-frequency water quality monitoring data with large fluctuations. Hence, to process the anomaly observations from the real-time water quality monitoring streams we combined the neural network based regression model with wavelet decomposition algorithms as illustrated in Figure 5.

Wavelet decomposition algorithms are well-known methods for capturing features of time series both in time and frequency domains. By applying wavelet decomposition to the original signals, a wavelet family that is correlated with the signal can be created. After that, wavelet denoising can be applied to the original signal to eliminate high-frequency noise.

One commonly used idea in anomaly detection is based on predictive models (Hill and Minsker (2010)). In this approach, one step ahead prediction is generated by learning from the previous observations. Then, the upper and lower threshold of the valid observations are calculated. Outliers can then be removed based on the

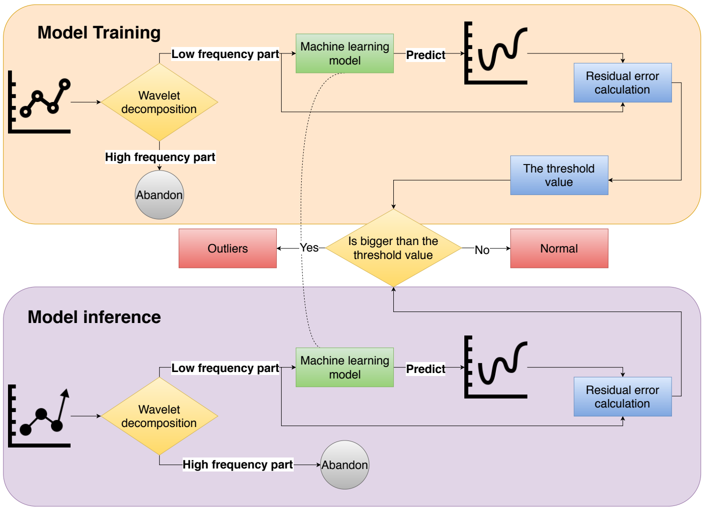


Figure 5. An outlier detection framework based on wavelet decomposition and neural network predictive model. The upper and lower block describe the model training and inference phase, respectively.

threshold determined. Thresholds are most often based on statistical analysis of the data, but may also come from historical data, experience or recommendation by a domain expert.

In the framework depicted in Figure 5, the water quality stream is first decomposed into a high-frequency and low-frequency component. The low-frequency signal is used to train a neural network-based predictive model while the high-frequency component is treated as noise. After obtaining a data-driven model with high predictive accuracy, the threshold for outlier detection is calculated based on the residual error between observations and predictions. In the inference phase, the well-trained data-driven model is applied to the target data streams. At each time index, if the water quality observations and the predicted water quality value have the residual errors higher than the threshold, the water quality observation at this time index is labelled as an outlier.

Real-time nitrate data collected between 12/2016 and 8/2018 from the Mulgrave River (GBR) has been used in testing our outlier detection framework. In the experiment, the nitrate data stream was first decomposed into a high-frequency and low-frequency component. After that, an ANN model was designed to predict the low-frequency signal and tread the high-frequency to be noise. The residuals of the true observations and model estimations over the training region were calculated. The threshold was then taken to be the mean of these residuals plus half a standard deviation. The results demonstrated that decomposition highlighted more of the outliers in general. All components of the high-frequency signal were classified as outliers, which also smoothed the sensor data.

3 CONCLUSIONS

Water quality modelling is a valuable tool to investigate, describe and predict the ecological state of the aquatic ecosystem. High-frequency water quality monitoring systems provide a vast amount of water quality observations, but these will require techniques to improve data streams and to predict trends in the data. In this paper, we illustrated various machine learning based modelling techniques investigated in the DigiscapeGBR Project (CSIRO (2019)) for solving three water quality challenges. Experimental results in different aquatic ecosystems demonstrate the efficiency in applying machine learning in the field of water quality prediction, imputation and outlier detection.

ACKNOWLEDGEMENT

We would like to thank the Great Barrier Reef Catchment Loads Monitoring Program (QLD (2019)) for providing valuable real-time water quality monitoring data sets.

References

Chang, F.-J., Y.-H. Tsai, P.-A. Chen, A. Coynel, and G. Vachaud (2015). Modeling water quality in an urban river using hydrological factors – data driven approaches. *Journal of Environmental Management 151*, 87–96.

CSIRO (2019). DigiscapeGBR. https://research.csiro.au/digiscape/digiscapes-projects/great-barrier-reef-and-sugarcane-production/. Accessed: 2019-04-20.

Dabrowski, J. J., A. Rahman, A. George, S. Arnold, and J. McCulloch (2018). State space models for forecasting water quality variables: an application in aquaculture prawn farming. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 177–185.

Di Blasi, J. P., J. M. Torres, P. G. Nieto, J. A. Fernández, C. D. Muñiz, and J. Taboada (2013). Analysis and detection of outliers in water quality parameters from different automated monitoring stations in the miño river basin (nw spain). *Ecological engineering 60*, 60–66.

Hill, D. J. and B. S. Minsker (2010). Anomaly detection in streaming environmental sensor data: A data-driven modeling approach. *Environmental Modelling & Software 25*(9), 1014–1022.

Hinton, G. E., N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Lea, C., M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager (2017). Temporal convolutional networks for action segmentation and detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1003–1012. IEEE.

QLD (2019). Qldmonitoring. https://water-monitoring.information.qld.gov.au. Accessed: 2019-04-20.

Ruben, G. B., K. Zhang, H. Bao, and X. Ma (2018). Application and sensitivity analysis of artificial neural network for prediction of chemical oxygen demand. *Water resources management 32*(1), 273–283.

Sarkar, A. and P. Pandey (2015). River water quality modelling using artificial neural network technique. *Aquatic Procedia 4*, 1070–1077.

Thorburn, P. J., P. Fitch, Y. Zhang, Y. Shendryk, T. Webster, J. Biggs, M. Mooij, C. Ticehurst, M. P. Vilas, and S. Fielke (2019). Helping farmers mitigate nutrient losses to the Great Barrier Reef through Digital Agriculture. In *Occasional Report, Fertiliser and Lime Research Centre, Massey University*, Volume 32, pp. 6.

Thorburn, P. J. and S. Wilkinson (2013). Conceptual frameworks for estimating the water quality benefits of improved agricultural management practices in large catchments. *Agriculture, Ecosystems & Environment 180*, 192 – 209.

Van Den Oord, A., S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu (2016). Wavenet: A generative model for raw audio. *CoRR abs/1609.03499*.

Vilas, M. P., C. L. Marti, M. P. Adams, C. E. Oldham, and M. R. Hipsey (2017). Invasive macrophytes control the spatial and temporal patterns of temperature and dissolved oxygen in a shallow lake: A proposed feedback mechanism of macrophyte loss. *Frontiers in Plant Science 8*, 2097.

Zhang, Y., P. Fitch, M. P. Vilas, and P. J. Thorburn (2019). Applying multi-layer artificial neural network and mutual information to the prediction of trends in dissolved oxygen. *Frontiers in Environmental Science 7*, 46.

Zhang, Y., P. J. Thorburn, and P. Fitch (2019). Multi-task temporal convolutional network for predicting water quality sensor data. In *26th International Conference on Neural Information Processing*. submitted.

Zhang, Y., P. J. Thorburn, X. Wei, and P. Fitch (2019). SSIM -a deep learning approach for recovering missing time series sensor data. *IEEE Internet of Things Journal 6*(4), 6618–6628.