H2Open Journal

IWA PUBLISHING

H2Open Journal Vol 3 No 1
doi: 10.2166/h2oj.2020.035

Quantifying natural organic matter concentration in water from climatological parameters using different machine learning algorithms

Sina Moradi[a,b], Anthony Agostino[a], Ziba Gandomkar[c], Seokhyeon Kim[d], Lisa Hamilton[e], Ashish Sharma[d], Rita Henderson[a] and Greg Leslie[b,*]

[a] Algae & Organic Matter Laboratory, School of Chemical Engineering, University of New South Wales, Sydney 2052, Australia

[b] UNESCO Centre for Membrane Science & Technology, School of Chemical Engineering, University of New South Wales, Sydney 2052, Australia

[c] Discipline of Medical Imaging Sciences, Faculty of Medicine and Health, University of Sydney, Sydney 2006, Australia

[d] School of Civil and Environmental Engineering, University of New South Wales, Sydney 2052, Australia

[e] Water and Catchment Protection, WaterNSW, Sydney 2150, Australia

*Corresponding author. E-mail: g.leslie@unsw.edu.au

Abstract

The present understanding of how changes in climate conditions will impact the flux of natural organic matter (NOM) from the terrestrial to aquatic environments and thus aquatic dissolved organic carbon (DOC) concentrations is limited. In this study, three machine learning algorithms were used to predict variations in DOC concentrations in an Australian drinking water catchment as a function of climate, catchment and physical water quality data. Four independent variables including precipitation, temperature, leaf area index and turbidity ($n = 5{,}540$) were selected from a large dataset to develop and train each machine learning model. The accuracy of the multivariable linear regression, support vector regression (SVR) and Gaussian process regression algorithms with different kernel functions was determined using adjusted R-squared (adj. R^2), root-mean-squared error (RMSE) and mean absolute error (MAE). Model accuracy was very sensitive to the time interval used to average climate observations prior to pairing with DOC observations. The SVR model with a quadratic kernel function and a 12-day time interval between climate and water quality observations outperformed the other machine learning algorithms (adj. $R^2 = 0.71$, RMSE = 1.9, MAE = 1.35). The area under the receiver operating characteristic curve method (AUC) confirmed that the SVR model could predict 92% of the elevated DOC observations; however, it was not possible to estimate DOC values at specific sampling sites in the catchment, probably due to the complex local geological and hydrological changes in the sites that directly surround and feed each sampling point. Further research is required to establish potential relationships between climatological data and NOM concentration in other water catchments – especially in the face of a changing climate.

Key words: climate, machine learning, natural organic matter, water quality

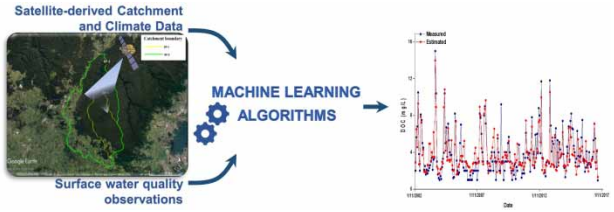
Highlights

- The application of different machine learning algorithms to quantify NOM in an Australian catchment.
- Investigate potential relationships between climatological factors and water quality parameters.

H2Open Journal Vol 3 No 1
doi: 10.2166/h2oj.2020.035

- The estimation of DOC concentration of water from simple measurable water quality and climatological variables.
- The detection of high DOC concentration events using the machine learning technique.

Graphical Abstract



INTRODUCTION

A warming climate is driving a suite of changes in global surface waters, from nutrient, temperature and sediment profiles in Chesapeake Bay, USA (Najjar *et al.* 2010) to the timing and intensity of monsoons in Southeast Asia (Loo *et al.* 2015). Higher temperatures impact water quality in lakes by reducing oxygen concentrations, increasing the release of phosphorus from sediments (Arnell *et al.* 2015) and increasing the concentration of dissolved organic matter (El-Jabi *et al.* 2014). Intense, but less frequent rainfall events will increase sediment loads while decreasing water security in drinking water catchments. These changes will have profound effects on the potable water supplies of current and future generations (Hansen *et al.* 2013).

The character and composition of natural organic matter (NOM) in freshwater ecosystems also appears to respond to variations in climatic conditions (Hejzlar *et al.* 2003; Delpla *et al.* 2009; Zhu *et al.* 2017; Gavin *et al.* 2018; Parr *et al.* 2019). This trend is also observed in Australian catchments with deleterious consequences for the performance of existing treatment infrastructure on more highly variable potable water supplies (Mohiuddin *et al.* 2014). Observed increases in dissolved organic carbon (DOC) concentrations have been linked to changes in climatic conditions (Tranvik & Jansson 2002; Freeman *et al.* 2004; Evans *et al.* 2006). Potential climatic drivers of the upward trends in DOC concentrations have included temperature (Evans *et al.* 2006), soil moisture (SM) (Hudson *et al.* 2003) and precipitation (Erlandsson *et al.* 2008). However, the present understanding of how changes in climatic conditions will continue to impact the flux of NOM from the terrestrial to aquatic environments and thus aquatic DOC concentrations is limited. Hence, establishing potential relationships between climatological data and NOM concentration in water catchment will be critical in planning for the delivery of potable water in a warmer climate.

The link between climate, catchment and water quality is dynamic, complex and with some exceptions not well understood. The expanded use of satellite-derived observations and increases in the type and frequency of water quality measurements has created more data but not necessarily better outcomes for the management of water quality. Consequently, there is an opportunity to apply advanced data-driven modelling techniques to catalogue and interrogate these complex datasets to identify the trends and interdependencies between the variables, and to perform dimensionality reduction of the variables (Lary *et al.* 2016; Ruescas *et al.* 2018). Supervised data-driven machine learning algorithms can potentially map DOC concentration in a catchment to climatological data without attempting to accurately model underlying processes. Machine learning methods have been widely used and achieved promising results in different environmental settings (Kim *et al.* 2014; Park *et al.* 2015; Granata *et al.* 2017; Ruescas *et al.* 2017), with the dataset size varying depending on the type of observation from $n =$ 63 (Kim *et al.* 2014) to $n = 2{,}255$ (Park *et al.* 2015) for analytical measurements and $n > 20{,}000$ for remote sensing applications (Ruescas *et al.* 2017). The application of data-driven modelling techniques

H2Open Journal Vol 3 No 1
doi: 10.2166/h2oj.2020.035

to estimate DOC concentrations has been demonstrated in other studies (Clair *et al.* 1999; Snauffer *et al.* 2018). For example, a neural network model was developed to estimate DOC concentration using hydrological, and climatic variables such as temperature, and total precipitation (Clair *et al.* 1999). In another application, the concentration of coloured dissolved organic matter was determined from remote sensing signals by comparing different machine learning regression approaches such as Gaussian process regression (GPR), support vector regression (SVR), random forest and kernel ridge regression (Ruescas *et al.* 2017, 2018).

The low computational cost and interpretability makes the multivariate linear regression (MLR) algorithm desirable as there are clear trade-offs between model complexity and interpretability. In many cases, particularly where limited data are available, more complex algorithms can achieve high levels of generalisation and prediction accuracy (Khalil *et al.* 2005). Among non-linear approaches, SVR is a nonparametric (i.e. not limited by a functional form) supervised learning algorithm that simultaneously can minimise prediction error and model complexity (Vapnik 2000). In contrast to linear and non-linear machine learning algorithms that are trained from exact values for every parameter in a function, GPR uses a (Gaussian) probability distribution over all possible values that fit the data. Exponential kernel function and squared exponential are widely used choices for covariance kernel functions (Jiang *et al.* 2019). GPR has been applied in regression problems in different environmental fields of studies including prediction of stream water temperature (Grbić *et al.* 2013), groundwater salinity (Lal & Datta 2018) or streamflow forecasting (Sun *et al.* 2014). However, the approach has not been used to explore correlations between DOC and catchment and climate variables.

The prospective application of an appropriate machine learning algorithm to a sufficiently large dataset may enable water utilities to predict what climate conditions are conducive to excursions in NOM in water catchments. The first challenge is to select the appropriate machine learning algorithm and develop a sufficiently large dataset to train and evaluate the performance of each machine learning algorithm. In most cases, the data required can be sourced from routinely collected water quality data, which are often fragmented by water utilities and catchment management authorities and temporally disconnected. An ever-growing library of climatological data is freely available from multiple weather and geographic satellite sources. However, as surface water is influenced by climate conditions through the transfer of momentum and matter, a time gap exists between climatological and water quality data. Moreover, as the catchment area increases, the lag time (here the optimal lagged day) increases (Rostami *et al.* 2018). Therefore, the second challenge is to explore the effects of different statistical pairing of observed DOC concentrations with averaged climatological variables from multiple time steps prior to the water quality observation. Thus, by combining and harmonising temporal differences between fragmented datasets, and training and validating an appropriate machine learning algorithm, it may be possible to identify potential relationships between the climatological and water quality data.

The objective of this paper was to evaluate three commonly used machine learning algorithms, namely MLR, SVR and GPR for the estimation of DOC concentration of water from simple measurable water quality and climatological variables using datasets from an Australian catchment. The secondary aim of this study was to establish new knowledge about the potential relationships between climatological factors and water quality parameters by using the exhaustive feature selection technique. This may enable water utilities to minimise the impact of changes in climate conditions on treatment plant performance and potable water quality.

METHODOLOGY

Study area

This study was conducted in the Nepean catchment in Australia (386 km^2) over a spatial domain of 33.81°S to 35.09°S and 149.97°E to 151.16°E with an elevation of 130–720 m above sea level

H2Open Journal Vol 3 No 1
doi: 10.2166/h2oj.2020.035

(Figure 1), The dominant climate class is temperate, characterised by no dry season, a warm summer and 800–1,600 mm of annual rainfall. The major land cover are trees (82%) with the remaining areas used for rural residential and agricultural purposes, including pasture and cropping (Peel *et al.* 2007b; Sixsmith *et al.* 2015; Office of Environment and Heritage 2017).

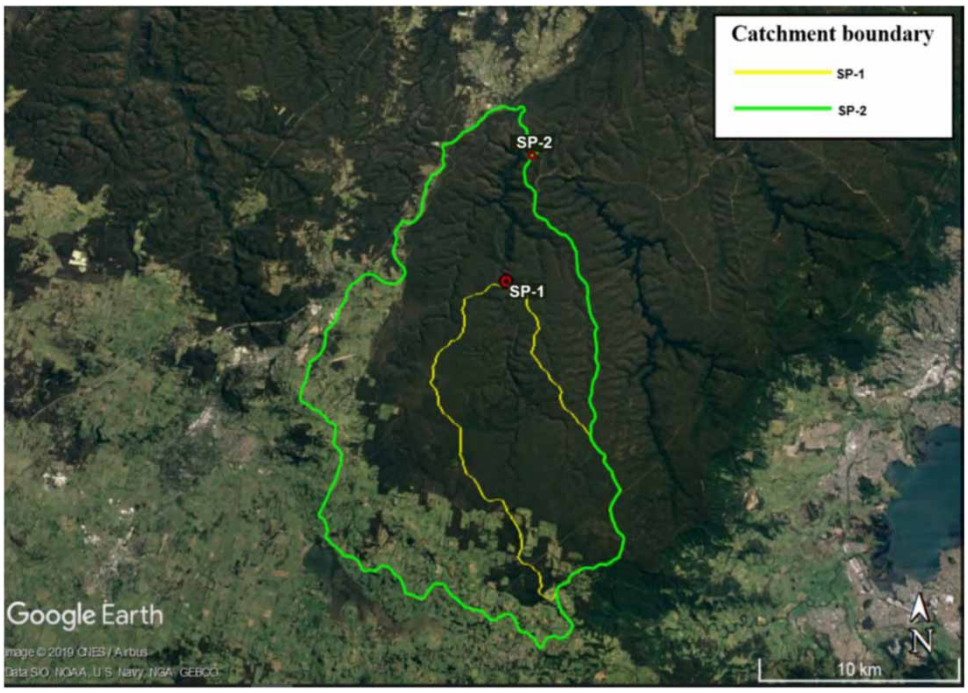


Figure 1 | Nepean dam catchment and two sampling points (SP-1 and SP-2).

Dataset

The dataset ($n = 5{,}540$) was compiled from surface water quality observations and climatological and catchment variables derived from satellites, and a land surface model with a temporal range of 15 years and 2 months (1 November 2002 to 1 January 2018) (Table S1, Supporting Information). A pre-processing step was employed to remove inaccuracies, including reading errors, abnormal data and missing values.

Water quality data

Water quality data were collected at the Burke River (SP-1), one of three main tributaries to Lake Nepean and Lake Nepean (SP-2) in the Nepean catchment (Figure S1, Supporting Information). Parameters obtained from routine water quality monitoring included pH, water temperature (°C), turbidity (NTU), alkalinity (mg $CaCO_3$/L), true colour (UTC) and DOC concentration (mg/L).

Catchment and climate data

The catchment boundary (DEM) was obtained from the Shuttle Radar Topography Mission (SRTM) and the Hydrologically Enforced Digital Elevation Model (DEM-H) Version 1.0 (Read *et al.* 2011); land cover (LC) was obtained from the Dynamic Land Cover Dataset Version 2.1 (Sixsmith *et al.* 2015) in 2012–2013, consisting of 22 LC classes; and climate class (CZ) was obtained from the updated Köppen–Geiger climate classification (Peel *et al.* 2007a) (Table S1). Daily rainfall (mm/day) and air temperature (°C) were obtained from the Australian Water Availability Project

H2Open Journal Vol 3 No 1
doi: 10.2166/h2oj.2020.035

(AWAP) dataset, gridded at 0.05° (Raupach *et al.* 2009). Mean temperatures were expressed as arithmetic averages of daily maximum and minimum temperatures.

Daily SM (%) and actual evapotranspiration (ET) (mm/day) were extracted from the Australian Water Resource Assessment Landscape (AWRA-L) dataset, gridded at 0.05° (Frost *et al.* 2016). Leaf area index (LAI) (m^2/m^2) defined as the one-sided green leaf area per unit ground surface area was calculated from 8-day composite Moderate Resolution Imaging Spectroradiometer (MODIS) measurements from Terra and Aqua satellites products (Didan *et al.* 2015). Raw LAI data were re-projected and resampled to 0.05° × 0.05° grids (identical to the gridded rainfall data) by using the MODIS re-projection tool (Dwyer & Schmidt 2006). Outliers were removed using the Savitzky–Golay filter (Savitzky & Golay 1964) in the TIMESAT software package (Jönsson & Eklundh 2004). The filtered 8-day LAI maps were linearly interpolated to a daily time scale for consistency with the other parameters in the dataset.

Data processing

Water quality data were collected on different time steps, and the dates of sample collection were not consistent. When multiple samples were taken within a day, the corresponding measurements were averaged to produce a single representative daily value. The climatological data series was complete and void of missing values, so additional data processing was not necessary. The climatological and water quality data series were merged with each other such that only the dates with both climatological and water quality data available could be considered for further data processing.

Establishing the machine learning algorithms

The workflow to establish, train and assess the machine learning algorithms is presented in Figure 2. To evaluate predictive models, the original dataset was randomly partitioned into 10 equal size non-overlapping subsets to enable the 10-fold cross-validation technique. Nine subsets were used for

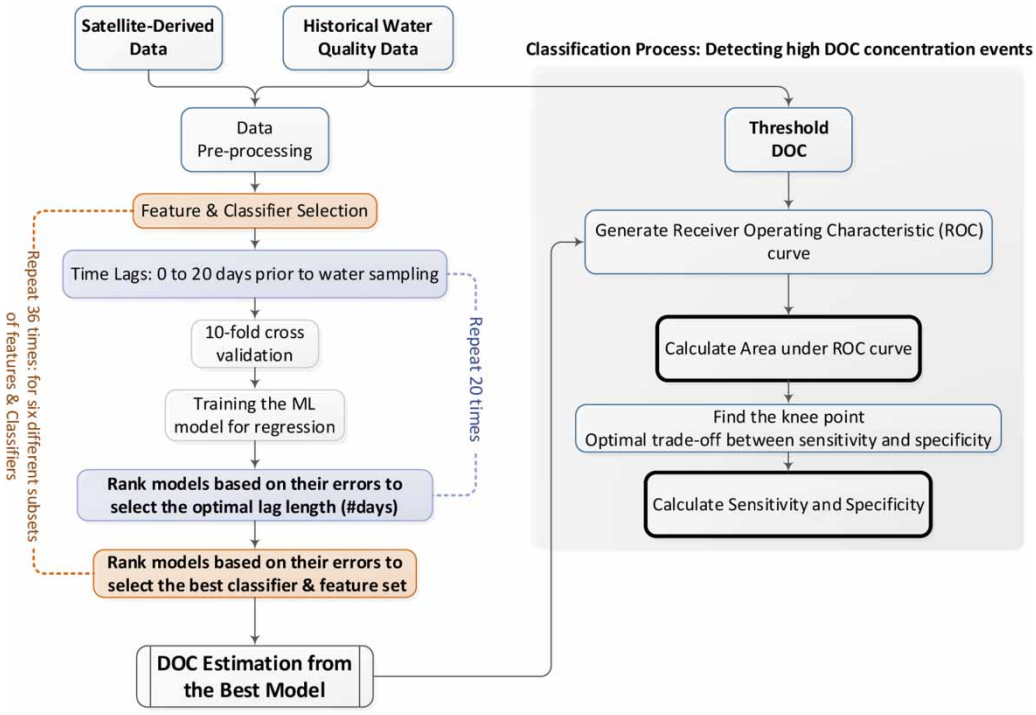


Figure 2 | The overall machine learning framework to establish, train and assess the machine learning algorithms.

H2Open Journal Vol 3 No 1
doi: 10.2166/h2oj.2020.035

training and one set for validation. The cross-validation process was repeated 10 times such that each of the 10 subsets was used exactly once as the validation data. Simple regression analysis was used to identify the potential correlations between the observed DOC concentrations (Target) and other water quality, catchment and climate data (Predictors). After understanding the associations between the predictors and the target, the three algorithms were programmed, trained and validated using MATLAB 2018b (Mathworks, Inc.).

Algorithm 1. MLR

The MLR model learns the linear function to map the climatological parameters to the DOC concentrations in the catchment. Equation (1) represents the relationship between the predicted DOC for the kth sample point, $\widehat{\mathrm{DOC}}_k$, using the MLR model:

$$\widehat{\mathrm{DOC}}_k = a_0 + \sum_{n=1}^{M} (a_n x_i) \tag{1}$$

where a_0 is the intercept and a linear term for each predictor (x_i). The MLR is the simplest algorithm and was used to establish the baseline response between target and predictors. The ordinary least square approximation by QR decomposition was used to estimate the algorithm parameters.

Algorithm 2. SVR

The SVR can be expressed as Equation (2):

$$\widehat{\mathrm{DOC}}_k = a_0 + \sum_{i=1}^{M} (a_i - a_i^*) \cdot K(x_i, x) \tag{2}$$

where K is the kernel function which transforms the data to map into a high-dimensional space. The effectiveness of several kernel functions, including linear, quadratic and cubic, to estimate DOC concentration using climatological parameters were tested. MATLAB allows the setting of certain parameters along with the kernel function as the SVR hyper parameters. Default MATLAB values were used for building the model. Briefly, the epsilon parameter, which represents the half width of epsilon-insensitive band and set to the interquartile range of DOC divided by 13.49 (i.e. an estimator for 10% of the standard deviation). When training an SVR, no penalty is associated with the training points inside the epsilon-insensitive band. The box constraint parameter, which controls the penalty of misclassified training samples, hence limiting model complexity, was set to a multiple of 10 times the epsilon value. Prior to training the model, the predictors were standardised, and the kernel scale parameter was set automatically by MATLAB.

Algorithm 3. GPR

The exponential kernel function and the squared exponential kernel function were used in this study, as they are widely used choices for covariance kernel functions (Jiang *et al.* 2019). The squared exponential covariance function used in the GPR algorithm was expressed as Equation (3):

$$k(x, x') = \sigma^2 \times e^{\left[\frac{-(x-x')^2}{2l^2}\right]} \tag{3}$$

where σ is the noise standard deviation and l is the characteristic length-scale which controls the effect of distance between x and x'. Intuitively, when x and x' are very close, then k (covariance function) leads to a higher value, which means $f(x)$ is highly correlated with $f(x')$. On the other hand, for two distant points, $k(x,x')$ is near to zero. This property of the squared exponential function is also highly desirable as for predicting the target variable at a new x, negligible effects from distant observations are anticipated.

H2Open Journal Vol 3 No 1
doi: 10.2166/h2oj.2020.035

Time-lag function

As the system time lag was unknown, the performance of the machine learning algorithms as measured by different statistical criteria (highlighted in this section) was evaluated as a function of the lag (T_D). Equation (4), where $\hat{y}$ and θ indicate the predicted value and model parameters, respectively, summarises the dominant delay estimation process. $Perf(\hat{y}, y)$ shows a performance metric describing the accuracy of the model. As implied by Equation (4), the largest spike (arg max) in the performance metric specified the dominant lag time.

$$\arg\max_{T_D} Perf(\hat{y}, y) \text{ where } \hat{y} = f(x_i, \theta) \tag{4}$$

Performance and accuracy assessment of the DOC models

Three summary statistics describing the accuracy of the models were used to assess the models' performances: adjusted R-squared (adj. R^2), the root-mean-squared error (RMSE) and mean absolute error (MAE). The adj. R^2 indicates the proportion of the total sum of squares explained by each model adjusted for the number of coefficients. To quantify the level of statistical significance in the performance of selected machine learning algorithms, the paired t-test as a commonly used statistical hypothesis test for comparing machine learning algorithms was applied into the 10-fold cross-validation results. The paired t-test results that are often a test statistic and a p-value can be interpreted to confirm whether there is a real or statistically significant difference between the machine learning algorithms (Nadeau & Bengio 2003; Bouckaert & Frank 2004). Although predicting the DOC as a continuous variable is highly desirable, ultimately predicting the events with high DOC concentrations might be more useful in practice. To do so, thresholding the predicted DOC value to produce a binary outcome was required.

To visualise and quantify the trade-off between the false alarms and missed events, the receiver operating characteristic (ROC) curve was generated for the best performing model (the classification process in Figure 2). The ROC curve plots true positive rate (that is known as sensitivity) against false positive rate (that is also known as one minus specificity), obtained by varying the probability threshold on the predicted output (DOC concentration) (Fawcett 2006). The sensitivity measures the accuracy of the model in predicting high DOC concentration events, while the specificity measures the accuracy of the model in predicting instances with below-threshold DOC concentration. The probability threshold was used to find out how accurately the model could determine if there exists a high DOC event or not. Since increases by as much as 100% in DOC concentrations were observed during wet weather events in other researches (Hinton *et al.* 1997; Schoenheinz & Grischek 2011; Whitworth *et al.* 2012; Dhillon & Inamdar 2013), a high DOC event was defined in this study when the DOC concentration was two times higher than the averaged observed historical DOC concentration in the dataset. To provide an aggregate measure to quantify the model performance across all possible threshold values for $\widehat{DOC}$, the area under the ROC curve (AUC) was calculated. The AUC value of 1.0 indicates a perfect model performance, while a value of 0.5 represents the chance-level (Kordestani *et al.* 2019).

RESULTS AND DISCUSSION

Removal of irrelevant data

Table 1 enumerates the basic statistics of the monitored water quality data at selected sampling points (SP-1 and SP-2 in Figure 1) including the mean and standard deviation of each monitored water

H2Open Journal Vol 3 No 1
doi: 10.2166/h2oj.2020.035

Table 1 | Summary of the basic statistical assessments of selected measured water quality variables in selected sampling sites

	SP-1			SP-2		
Variable	**Mean**	**Standard deviation**	**R^2 with DOC**	**Mean**	**Standard deviation**	**R^2 with DOC**
pH	6.53	0.50	0.00	7.31	0.38	0.00
Temperature (°C)	16.00	5.30	0.26	14.81	3.52	0.17
Turbidity (NTU)	7.72	4.94	0.30	2.15	1.86	0.10
Alkalinity (mg $CaCo_5$/L)	5.17	3.54	0.00	11.07	1.08	0.00
True colour (HU)	23.12	10.40	0.22	7.14	1.86	0.25
DOC concentration (mg/L)	4.30	1.81	1.00	3.7	0.99	1.00

quality data and their R^2 with DOC concentration for each site. pH and alkalinity were dropped from the analysis, as they were variables exhibiting an insignificant relationship with DOC concentration. It should be noted that some studies suggested rising DOC concentrations may have been linked to increases in alkalinity or pH, associated with recovery from acidification (Evans *et al.* 2005; Winterdahl *et al.* 2014). However, there are some studies that challenged the role of pH (Couture *et al.* 2012; Pärn & Mander 2012). Although the temporal correlations were not strong between pH and DOC concentrations in this dataset, the possibility that as pH increases the DOC concentrations increase in this catchment is not dismissed. True colour was infrequently measured in our dataset ($n = 1{,}758$ compared with $n = 10{,}761$ for turbidity at SP-1), thus it was excluded as a parameter in DOC modelling. Temperature and turbidity were retained as water quality variables for the modelling of the DOC concentration.

Selection of correct predictor variables

Climatological data including precipitation (Pre.), temperature (Temp.), SM, ET and LAI measured at SP-1 from November 2002 to January 2018 are shown in Figure S2. To select the correct climatological variables for DOC estimation, Pearson's correlation coefficient of each climatological variable and selected water quality parameters with DOC concentration in selected sampling sites were determined (Table 2 and Table S2). Strong correlations were observed between SM and precipitation data ($r = 0.91$ at SP-1) and between temperature and ET data ($r = 0.80$ at SP-1) (Table 2 and Table S2). To ensure parameters used to estimate DOC concentration were independent from each other, parameters with high correlation were excluded. Therefore, two pathways arise for potential machine learning model input variables, one considering precipitation and temperature and the other including SM and ET. The effects of including more input variables on estimated DOC concentration were also studied by increasing the number of model input variables from only two variables

Table 2 | Pearson's correlation coefficient of climatological variables and selected water quality parameters with DOC concentration of SP-1 dataset

MP1	**DOC**	**Turb.**	**Pre.**	**Temp.**	**SM**	**ET**	**LAI**
DOC	1.00						
Turb.	0.29	1.00					
Pre.	0.45	0.19	1.00				
Temp.	0.26	0.06	0.07	1.00			
SM	0.41	0.18	0.91	0.07	1.00		
ET	0.29	0.02	0.34	0.80	0.38	1.00	
LAI	0.12	0.00	0.11	0.15	0.05	0.18	1.00

H2Open Journal Vol 3 No 1
doi: 10.2166/h2oj.2020.035

(Pre. and Temp. versus SM and ET) to three and four input variables by adding turbidity and LAI, respectively.

Impact of considering time lags for the climatological data on estimated DOC concentrations

To investigate the effects of considering time lags for the climatological data on estimated DOC concentrations, time lags of up to 20 days prior to water sampling for the climatological data were investigated in this study. Adj. R^2 and RMSE values of estimated DOC concentrations at SP-1 using MLR as well as SVR and GPR with different kernel functions are presented in Figure 3(a)–3(f) and Figure S3(a)–S3(f),

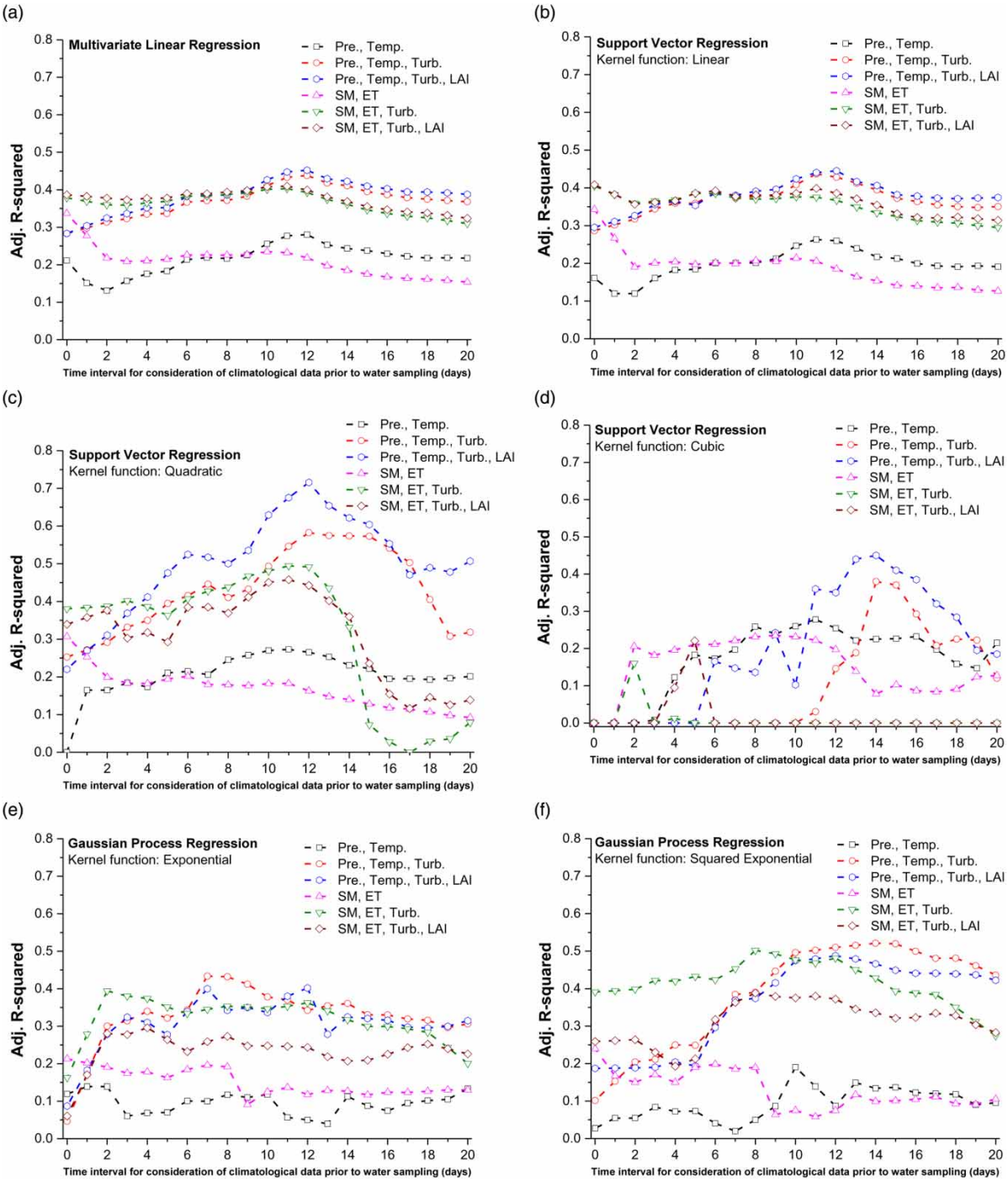


Figure 3 | Adj. R^2 values of estimated DOC concentration at SP-1 using MLR, SVR and GPR with different kernel functions.

H2Open Journal Vol 3 No 1
doi: 10.2166/h2oj.2020.035

respectively. Figure S4(a)–S4(f) represents how MAE values of estimated DOC concentration at SP-1 varied by increasing the time interval for consideration of climatological data prior to water sampling. The results in Figure 3(a)–3(f) showed that increasing the number of inputs gave better results, especially with the more sophisticated machine learning approaches. The model performance was improved by considering turbidity and LAI as input variables, and the first scenario (Pre. and Temp.) resulted better than the second one (SM and ET) (Figure 3(a)–3(f)). SVR with a linear kernel function (Figure 3(b)) resulted in slightly better performance than MLR in adj. R^2 values when the time lag between climatological data and water quality data was increased. Changing the kernel function from linear to quadratic also improved the model estimation results (adj. R^2 of the model was increased) (Figure 3(c)). However, applying SVR with a cubic kernel function was unsuccessful for this dataset for DOC concentration estimation (Figure 3(d)). Similarly, GPR with kernel functions either exponential (Figure 3(e)) or squared exponential (Figure 3(f)) exhibited poor performance with maximum adj. R^2 values on average less than 0.5. The adj. R^2 increased when the time lag between the climatological data and the day of water sampling was increased from 1 to 12 days, beyond which it continued to decrease (Figure 3(a)–3(d)).

Therefore, using SVR with a quadratic kernel function with Pre., Temp., LAI and Turb. as model input variables with a 12-day time lag between climatological and water quality data outperformed the other selected machine learning algorithms in this study in terms of statistical indices (adj. $R^2 = 0.71$, RMSE = 1.9, MAE = 1.35). To estimate whether the differences between the performance of SVR with a quadratic kernel function and other selected machine learning algorithms are true and reliable or are just due to statistical chance, a paired *t*-test was conducted on 10-fold cross-validation results, while a 12-day time lag between climatological and water quality data was considered. The paired *t*-test results between SVR with a quadratic kernel function and other machine learning algorithms rejected the null hypothesis at the 5% significance level, and the averaged *p*-value was 1.37% which is smaller than the considered significance level (i.e. 5%). Hence, the results statistically provided convincing evidence that machine learning algorithms performed differently and any observed difference in the performance of SVR with a quadratic kernel function is likely due to a difference in the models. Therefore, the SVR showed a better capability than MLR and GPR to handle the nonlinearity and complexity of climatological characteristics of the catchment. However, it should be acknowledged that the performance of the GPR model was investigated with only two kernel functions, the exponential and the squared exponential. Exploring other kernels could be a potential future work to this study and might result in a higher performance metrics for the GPR.

The trend of RMSE (Figure S3(a)–S3(f)) was opposite to that of adj. R^2, as the objective function with RMSE was to minimise the estimation error. It should be noted that by using the entire dataset in the hyperparameter selection process, a bias is introduced, which likely results in an overly optimistic performance metrics. The hyperparameters were the best model types (i.e. SVR with quadratic kernel function), the best set of features (Pre., Temp., LAI and Turb.) and the most appropriate value time delay (i.e. 12 days). It should be noted that the 12-day time interval would not be a universal value for all the other catchments, as the time interval could be related to some physical characteristics of the catchment such as catchment area/slope or the LC.

Machine learning model performance assessment

One potential avenue is to explore the performance of the best model for a completely blinded dataset, which was not used for the feature and model selection. Therefore, a 12-day time lag was chosen to consider the impacts of climatological data prior to water sampling, as it led to the highest adj. R^2 value and lowest RMSE value in this study (Figure 3(c)). The measured and estimated DOC concentrations and their scatter plot using the SVR model with the quadratic kernel function at SP-1 are presented in Figure 4. As seen from the figure, the plotted data points revealed a good agreement

H2Open Journal Vol 3 No 1
doi: 10.2166/h2oj.2020.035

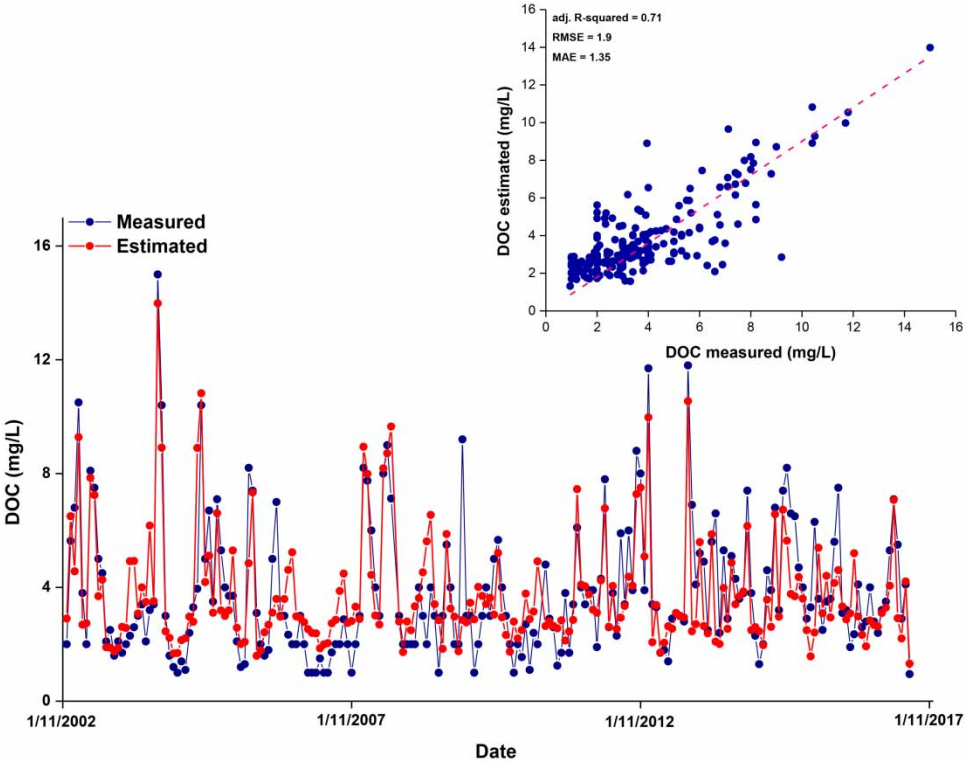


Figure 4 | The measured and estimated DOC concentrations and the scatter plot using the SVR model with the quadratic kernel function at SP-1.

between measured and estimated DOC concentrations, as they generally correlated close towards the 1:1 sloped line.

ROC analysis

To investigate if the developed SVR model with a quadratic kernel function and a 12-day time lag between climatological and water quality data could detect high DOC concentration events, the ROC analysis was carried out. The averaged measured DOC concentration at SP-1 from November 2002 to January 2018 was 3.4 mg/L (Figure 4). In this study, the DOC concentration threshold for such events was set to be two times higher than the averaged observed DOC concentrations over the past 15 years (6.8 mg/L at SP-1). Figure 5 shows the ROC curve for the high DOC concentration events using the SVR model with the quadratic kernel function with Pre., Temp., LAI and Turb. as model input variables and a 12-day time lag between climatological and water quality data at SP-1. Different cut-off values can be selected from the ROC curve. Here, as high sensitivity will guarantee that most of the high DOC concentration events will be detected, a lax operating point is desirable. Therefore, a knee point in the ROC curve, which ensures a sensitivity of 90%, was selected as the operating point. The corresponding sensitivity and specificity were 0.89 and 0.88, respectively (denoted by a red circle in Figure 5). The confidence intervals in each point were computed by generating 100 bootstrap replicas. It was found that the AUC value for the SVR model with the quadratic kernel function and a 12-day time lag was 0.92 which is above the random level (i.e. 0.5), indicating that the developed SVR model, which uses climatological variables as model inputs, is capable of being used as an alarm system, indicating a possible high DOC event.

Adj. R^2, RMSE and MAE values for the estimated DOC concentrations at SP-2 with the same machine learning algorithms are shown in Figure S5(a)–S5(f), Figure S6(a)–S6(f) and

H2Open Journal Vol 3 No 1
doi: 10.2166/h2oj.2020.035

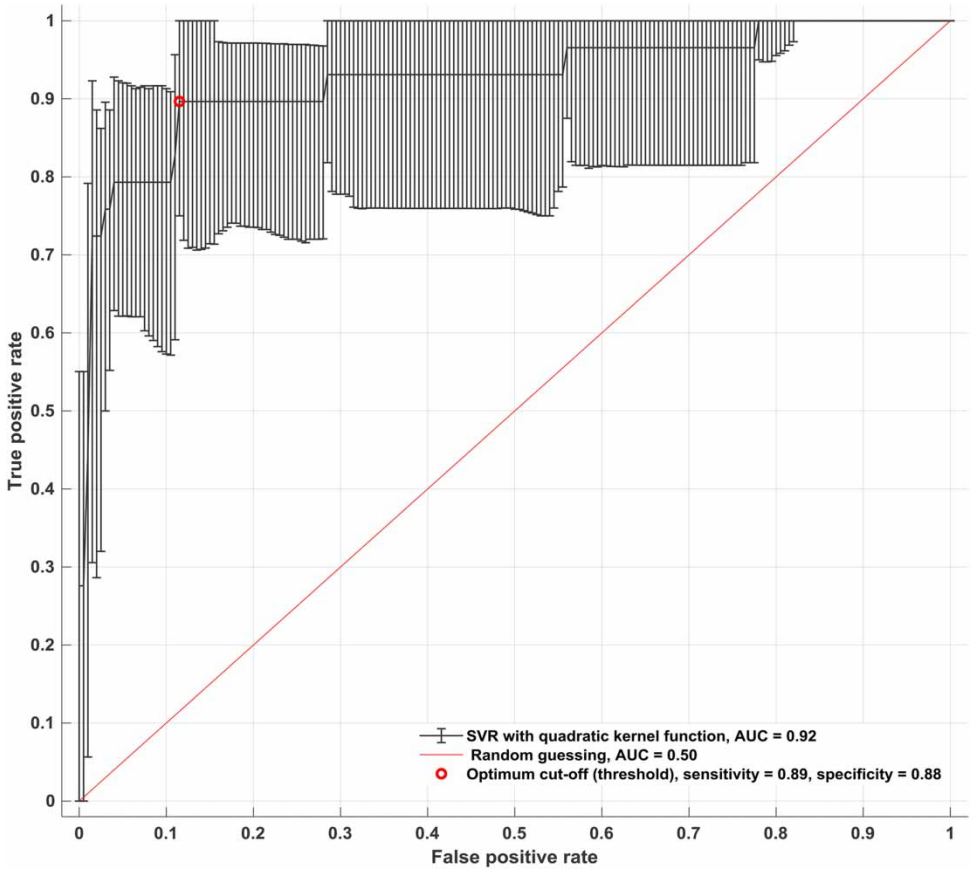


Figure 5 | The ROC curve and the AUC corresponding to the high DOC concentration events using the SVR model with the quadratic kernel function with Pre., Temp., LAI and Turb. as model input variables and a 12-day time lag between climatological and water quality data at SP-1.

Figure S7(a)–S7(f), respectively. Unlike SP-1 that is a river-based sampling point, SP-2 is located at a lake whose water comes from three main tributaries (SP-1 is at one of these three main tributaries into the lake). The results show that none of the applied machine learning algorithms could estimate the DOC concentration at SP-2 (adj. R^2 <0.3). As shown in Figure S5(a)–S5(f), Figure S6(a)–S6(f) and Figure S7(a)–S7(f), estimated DOC concentrations at SP-2 did not improve by including more input parameters to machine learning algorithms, or by increasing the time lag between climatological and water quality datasets. This could largely be due to the complex local geological and hydrological changes in the sites that directly surround and feed each sampling point or because the lake water combines water from its tributaries with potentially different physiochemical and climatological characteristics. One future approach to deal with lake-based sites could be to consider data from all input tributaries. Another possible reason, that would make it hard to understand how water quality parameters in a lake are affected by climatological parameters, could be the inherent variability in lake characteristics such as lake size, shape and depth compared to the rivers. It could also be due to the complex nature of lake–atmosphere interactions (O'Reilly *et al.* 2015; Winslow *et al.* 2015) or water clarity in the lake as it influences the depth range over which heat can be absorbed (Rose *et al.* 2016) or the strength of lake stratification (Winslow *et al.* 2015, 2017). Hence, the response of the lake to changing climatological parameters was complex, and considering only the climatological parameters could not estimate the DOC concentration at the lake sampling point in our study (SP-2).

H2Open Journal Vol 3 No 1
doi: 10.2166/h2oj.2020.035

CONCLUSION

This research utilised machine learning and satellite data to develop and train a model to estimate DOC concentration in water in an Australian catchment. Given the ease of accessing high-resolution satellite data, it is feasible to use climatological data derived from the satellite datasets and link them with water quality indicators to meet the current and future challenges in large-area water quality monitoring. The analysis of the results showed that precipitation, temperature, LAI and turbidity yielded the optimal results using the SVR model with the quadratic kernel function. Considering the impact of time lag on climatological data prior to water sampling showed an impact on model accuracy, and a 12-day time lag between climatological and water quality data brought far better results for the datasets used in this study in terms of statistical indices (adj. $R^2 = 0.71$, RMSE = 1.9, MAE = 1.35). To show the usefulness of the proposed method compared with other traditional model and kernel-based models, different machine learning models were constructed on the dataset. Experimental results show that the forecasting capability of the SVR model outperforms those of other kernel-based models, thereby generating more accurate results.

From the ROC analysis, the developed SVR model with the quadratic kernel function can be successfully used to indicate possible high DOC events (AUC = 92%). Although the results showed that the SVR model works well for the river-based sampling point, the response of the lake to changing climatological parameters was complex and none of the applied machine learning algorithms could estimate the DOC concentration at lake-based sampling point, probably due to higher water residence times in lakes than rivers or diverse variability in lake characteristics such as lake size, shape and depth. Further studies would be needed to test and clarify the effect of considering other potential lake characteristics as model input variables for the DOC estimation at lake-based sampling points and the potential for utilising the approach to river-based modelling used in this study as inputs to DOC estimation models for lake-based sites.

ACKNOWLEDGEMENTS

This research was supported via the Australian Research Council's Linkage Projects funding scheme (project number LP160100620) which included support from Sydney Water and WaterNSW. In addition, the authors thank Water Research Australia for the PhD top-up scholarship provided to Anthony Agostino. The authors also acknowledge the support of the UNESCO Centre for Membrane Science and Technology.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

REFERENCES

Arnell, N. W., Halliday, S. J., Battarbee, R. W., Skeffington, R. A. & Wade, A. J. 2015 The implications of climate change for the water environment in England. *Progress in Physical Geography: Earth and Environment* **39**, 93–120.

Bouckaert, R. R. & Frank, E. 2004 Evaluating the replicability of significance tests for comparing learning algorithms. In: *Advances in Knowledge Discovery and Data Mining* (Dai, H., Srikant, R. & Zhang, C., eds). Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 3–12.

Clair, T. A., Ehrman, J. M. & Higuchi, K. 1999 Changes in freshwater carbon exports from Canadian terrestrial basins to lakes and estuaries under a 2 × CO_2 atmospheric scenario. *Global Biogeochemical Cycles* **13**, 1091–1097.

Couture, S., Houle, D. & Gagnon, C. 2012 Increases of dissolved organic carbon in temperate and boreal lakes in Quebec, Canada. *Environmental Science and Pollution Research* **19**, 361–371.

H2Open Journal Vol 3 No 1
doi: 10.2166/h2oj.2020.035

Delpla, I., Jung, A. V., Baures, E., Clement, M. & Thomas, O. 2009 Impacts of climate change on surface water quality in relation to drinking water production. *Environment International* **35**, 1225–1233.

Dhillon, G. S. & Inamdar, S. 2013 Extreme storms and changes in particulate and dissolved organic carbon in runoff: entering uncharted waters? *Geophysical Research Letters* **40**, 1322–1327.

Didan, K., Barreto-Munoz, A., Solano, R. & Huete, A. 2015 *MODIS Vegetation Index User's Guide (MOD13 Series), Version 3.00, June 2015 (Collection 6)*. Vegetation Index and Phenology Lab, The University of Arizona, pp. 1–32.

Dwyer, M. J. & Schmidt, G. 2006 The MODIS reprojection tool. In: *Earth Science Satellite Remote Sensing Vol. 2: Data, Computational Processing, and Tools* (Qu, J. J., Gao, W., Kafatos, M., Murphy, R. E. & Salomonson, V. V., eds). Tsinghua University Press, Beijing and Springer-Verlag GmbH Berlin Heidelberg, pp. 162–177.

El-Jabi, N., Caissie, D. & Turkkan, N. 2014 Water quality index assessment under climate change. *Journal of Water Resource and Protection* **6**, 533–542.

Erlandsson, M., Buffam, I., Fölster, J., Laudon, H., Temnerud, J., Weyhenmeyer, G. A. & Bishop, K. 2008 Thirty-five years of synchrony in the organic matter concentrations of Swedish rivers explained by variation in flow and sulphate. *Global Change Biology* **14**, 1191–1198.

Evans, C. D., Monteith, D. T. & Cooper, D. M. 2005 Long-term increases in surface water dissolved organic carbon: observations, possible causes and environmental impacts. *Environmental Pollution* **137**, 55–71.

Evans, C. D., Chapman, P. J., Clark, J. M., Monteith, D. T. & Cresser, M. S. 2006 Alternative explanations for rising dissolved organic carbon export from organic soils. *Global Change Biology* **12**, 2044–2053.

Fawcett, T. 2006 An introduction to ROC analysis. *Pattern Recognition Letters* **27**, 861–874.

Freeman, C., Fenner, N., Ostle, N. J., Kang, H., Dowrick, D. J., Reynolds, B., Lock, M. A., Sleep, D., Hughes, S. & Hudson, J. 2004 Export of dissolved organic carbon from peatlands under elevated carbon dioxide levels. *Nature* **430**, 195–198.

Frost, A., Ramchurn, A. & Smith, A. 2016 *The Bureau's Operational AWRA Landscape (AWRA-L) Model*. Bureau of Meteorology, Melbourne, p. 47.

Gavin, A. L., Nelson, S. J., Klemmer, A. J., Fernandez, I. J., Strock, K. E. & McDowell, W. H. 2018 Acidification and climate linkages to increased dissolved organic carbon in high-elevation lakes. *Water Resources Research* **54**, 5376–5393.

Granata, F., Papirio, S., Esposito, G., Gargano, R. & De Marinis, G. 2017 Machine learning algorithms for the forecasting of wastewater quality indicators. *Water* **9**, 105.

Grbić, R., Kurtagić, D. & Slišković, D. 2013 Stream water temperature prediction based on Gaussian process regression. *Expert Systems with Applications* **40**, 7407–7414.

Hansen, J., Kharecha, P., Sato, M., Masson-Delmotte, V., Ackerman, F., Beerling, D. J., Hearty, P. J., Hoegh-Guldberg, O., Hsu, S.-L., Parmesan, C., Rockstrom, J., Rohling, E. J., Sachs, J., Smith, P., Steffen, K., Van Susteren, L., von Schuckmann, K. & Zachos, J. C. 2013 Assessing 'dangerous climate change': required reduction of carbon emissions to protect young people, future generations and nature. *PLoS One* **8**, e81648.

Hejzlar, J., Dubrovský, M., Buchtele, J. & Růžička, M. 2003 The apparent and potential effects of climate change on the inferred concentration of dissolved organic matter in a temperate stream (the Malše River, South Bohemia). *Science of the Total Environment* **310**, 143–152.

Hinton, M. J., Schiff, S. L. & English, M. C. 1997 The significance of storms for the concentration and export of dissolved organic carbon from two Precambrian Shield catchments. *Biogeochemistry* **36**, 67–88.

Hudson, J., Dillon, P. & Somers, K. 2003 Long-term patterns in dissolved organic carbon in boreal lakes: the role of incident radiation, precipitation, air temperature, southern oscillation and acid deposition. *Hydrology and Earth System Sciences* **7**, 390–398.

Jiang, S., Shen, X. & Zheng, Z. 2019 Gaussian process-based hybrid model for predicting oxygen consumption in the converter steelmaking process. *Processes* **7**, 352.

Jönsson, P. & Eklundh, L. 2004 TIMESAT – a program for analyzing time-series of satellite sensor data. *Computers & Geosciences* **30**, 833–845.

Khalil, A., Almasri, M. N., McKee, M. & Kaluarachchi, J. J. 2005 Applicability of statistical learning algorithms in groundwater quality modeling. *Water Resources Research* **41**, W05010–05026.

Kim, Y. H., Im, J., Ha, H. K., Choi, J.-K. & Ha, S. 2014 Machine learning approaches to coastal water quality monitoring using GOCI satellite data. *GIScience & Remote Sensing* **51**, 158–174.

Kordestani, M. D., Naghibi, S. A., Hashemi, H., Ahmadi, K., Kalantar, B. & Pradhan, B. 2019 Groundwater potential mapping using a novel data-mining ensemble model. *Hydrogeology Journal* **27**, 211–224.

Lal, A. & Datta, B. 2018 *Genetic Programming and Gaussian Process Regression Models for Groundwater Salinity Prediction: Machine Learning for Sustainable Water Resources Management*.

Lary, D. J., Alavi, A. H., Gandomi, A. H. & Walker, A. L. 2016 Machine learning in geosciences and remote sensing. *Geoscience Frontiers* **7**, 3–10.

Loo, Y. Y., Billa, L. & Singh, A. 2015 Effect of climate change on seasonal monsoon in Asia and its impact on the variability of monsoon rainfall in Southeast Asia. *Geoscience Frontiers* **6**, 817–823.

Mohiuddin, A., Rajanayagam, C. & Kearney, C. 2014 Optimisation of non-ionic polymer to address production issues with high-color low-turbidity raw water. *Water: Journal of the Australian Water Association* **41**, 58–63.

Nadeau, C. & Bengio, Y. 2003 Inference for the generalization error. *Machine Learning* **52**, 239–281.

Najjar, R. G., Pyke, C. R., Adams, M. B., Breitburg, D., Hershner, C., Kemp, M., Howarth, R., Mulholland, M. R., Paolisso, M., Secor, D., Sellner, K., Wardrop, D. & Wood, R. 2010 Potential climate-change impacts on the Chesapeake Bay. *Estuarine, Coastal and Shelf Science* **86**, 1–20.

H2Open Journal Vol 3 No 1
doi: 10.2166/h2oj.2020.035

Office of Environment and Heritage 2017 *Upper Nepean State Conservation Area Draft Plan of Management*. Office of Environment and Heritage, Sydney, Australia.

O'Reilly, C. M, Sharma, S., Gray, S., Hampton, D. K., Read, S. E., Rowley, J. S., Schneider, R. J., Lenters, P., McIntyre, J. D., Kraemer, P. B., Weyhenmeyer, B. M., Straile, G. A., Dong, D., Adrian, B., Allan, R., Anneville, M. G., Arvola, O., Austin, L., Bailey, J., Baron, J. L., Brookes, J. S., de Eyto, J. D., Dokulil, E., Hamilton, M. T., Havens, D. P., Hetherington, K., Higgins, A. L., Hook, S. N., Izmest'eva, S., Joehnk, L. R., Kangur, K. D., Kasprzak, K., Kumagai, P., Kuusisto, M., Leshkevich, E., Livingstone, G., MacIntyre, D. M., May, S., Melack, L., Mueller-Navarra, J. M., Naumenko, D. C., Noges, M., Noges, P., North, T., Plisnier, R. P., Rigosi, P.-D., Rimmer, A., Rogora, A., Rudstam, M., Rusak, J. A., Salmaso, N., Samal, N. R., Schindler, D. E., Schladow, G., Schmid, M., Schmidt, S. R., Silow, E., Soylu, E., Teubner, K., Verburg, P., Voutilainen, A., Watkinson, A., Williamson, C. E. & Zhang, G. 2015 Rapid and highly variable warming of lake surface waters around the globe. *Geophysical Research Letters* **42**, 10773–10781.

Park, Y., Cho, K. H., Park, J., Cha, S. M. & Kim, J. H. 2015 Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea. *Science of the Total Environment* **502**, 31–41.

Pärn, J. & Mander, Ü. 2012 Increased organic carbon concentrations in Estonian rivers in the period 1992–2007 as affected by deepening droughts. *Biogeochemistry* **108**, 351–358.

Parr, T. B., Inamdar, S. P. & Miller, M. J. 2019 Overlapping anthropogenic effects on hydrologic and seasonal trends in DOC in a surface water dependent water utility. *Water Research* **148**, 407–415.

Peel, M. C., Finlayson, B. L. & McMahon, T. A. 2007a Updated world map of the Köppen-Geiger climate classification. *Hydrology and Earth System Sciences Discussions* **4**, 439–473.

Peel, M. C., Finlayson, B. L. & McMahon, T. A. 2007b Updated world map of the Köppen-Geiger climate classification. *Hydrology and Earth System Sciences* **11**, 1633–1644.

Raupach, M., Briggs, P., Haverd, V., King, E., Paget, M. & Trudinger, C. 2009 *Australian Water Availability Project (AWAP): CSIRO Marine and Atmospheric Research Component: Final Report for Phase 3*. Centre for Australian Weather and Climate Research (Bureau of Meteorology and CSIRO), Melbourne, Australia, 67.

Read, A., Dowling, T., Gallant, J., Tickle, P. K. & Wilson, N. 2011 1Second SRTM Derived Hydrological Digital Elevation Model (DEM-H) version 1.0. Commonwealth of Australia (Geoscience Australia).

Rose, K. C., Winslow, L. A., Read, J. S. & Hansen, G. J. A. 2016 Climate-induced warming of lakes can be either amplified or suppressed by trends in water clarity. *Limnology and Oceanography Letters* **1**, 44–53.

Rostami, S., He, J. & Hassan, Q. K. 2018 Riverine water quality response to precipitation and its change. *Environments* **5**, 8.

Ruescas, A. B., Hieronymi, M., Koponen, S., Kallio, K. & Camps-Valls, G. 2017 Retrieval of coloured dissolved organic matter with machine learning methods. In: *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 2187–2190.

Ruescas, A. B., Hieronymi, M., Mateo-Garcia, G., Koponen, S., Kallio, K. & Camps-Valls, G. 2018 Machine learning regression approaches for colored dissolved organic matter (CDOM) retrieval with S2-MSI and S3-OLCI simulated data. *Remote Sensing* **10**, 1–20.

Savitzky, A. & Golay, M. J. E. 1964 Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry* **36**, 1627–1639.

Schoenheinz, D. & Grischek, T. 2011 Behavior of dissolved organic carbon during bank filtration under extreme climate conditions. In: *Riverbank Filtration for Water Security in Desert Countries* (Shamrukh, M., ed.). Springer Netherlands, Dordrecht, pp. 51–67.

Sixsmith, J., Thankappan, M., McIntyre, A., Tan, P. & Lymburner, L. 2015 *Dynamic Land Cover Dataset Version 2.1*. Geoscience Australia, Sydney, Australia.

Snauffer, A. M., Codden, C., Stubbins, A. & Mueller, A. V. 2018 High-frequency saltmarsh dissolved organic carbon estimates via machine learning. In: *AGU Fall Meeting Abstracts*, pp. EP51E-1875.

Sun, A. Y., Wang, D. & Xu, X. 2014 Monthly streamflow forecasting using Gaussian process regression. *Journal of Hydrology* **511**, 72–81.

Tranvik, L. J. & Jansson, M. 2002 Terrestrial export of organic carbon. *Nature* **415**, 861–862.

Vapnik, V. 2000 *The Nature of Statistical Learning Theory*. Springer, New York.

Whitworth, K. L., Baldwin, D. S. & Kerr, J. L. 2012 Drought, floods and water quality: drivers of a severe hypoxic blackwater event in a major river system (the southern Murray–Darling basin, Australia). *Journal of Hydrology* **450–451**, 190–198.

Winslow, L. A., Read, J. S., Hansen, G. J. A. & Hanson, P. C. 2015 Small lakes show muted climate change signal in deepwater temperatures. *Geophysical Research Letters* **42**, 355–361.

Winslow, L. A., Hansen, G. J. A., Read, J. S. & Notaro, M. 2017 Large-scale modeled contemporary and future water temperature estimates for 10774 Midwestern U.S. Lakes. *Scientific Data* **4**, 170053–170053.

Winterdahl, M., Bishop, K. & Erlandsson, M. 2014 *Acidification, Dissolved Organic Carbon (DOC) and Climate Change*, pp. 281–287.

Zhu, J., Hu, H., Tao, S., Chi, X., Li, P., Jiang, L., Ji, C., Zhu, J., Tang, Z., Pan, Y., Birdsey, R. A., He, X. & Fang, J. 2017 Carbon stocks and changes of dead organic matter in China's forests. *Nature Communications* **8**, 151.