





# A data mining approach to evaluate suitability of dissolved oxygen sensor observations for lake metabolism analysis

Kohji Muraoka <sup>1</sup>, Paul Hanson <sup>2</sup>, Eibe Frank <sup>3</sup>, Meilan Jiang,<sup>4</sup> Kenneth Chiu,<sup>5</sup> David Hamilton <sup>6\*</sup>

<sup>1</sup>Environmental Research Institute, University of Waikato, Hamilton, New Zealand

<sup>2</sup>Center for Limnology, University of Wisconsin-Madison, Madison, Wisconsin

<sup>3</sup>Department of Computer Science, University of Waikato, Hamilton, New Zealand

<sup>4</sup>Center for EcoInformatics, Konkuk University, Seoul, Republic of Korea

<sup>5</sup>Computer Science Department, Binghamton University, Binghamton, New York

<sup>6</sup>Australian Rivers Institute, Griffith University, Brisbane, Queensland, Australia

## Abstract

Despite rapid growth in continuous monitoring of dissolved oxygen for lake metabolism studies, the current best practice still relies on visual assessment and manual data filtering of sensor observations by experienced scientists in order to achieve meaningful results. This time consuming approach is fraught with potential for inconsistency and individual subjectivity. An automated method to assure the quality of data for the purpose of metabolism modeling is clearly needed to obtain consistent results representative of collective expertise. We used a hybrid approach of expert panel and data mining for data filtration. Symbolic Aggregate approXimation (SAX) treats discretized numerical timeseries segments as symbolic indications, creating a series of strings which are literally comparable to human words and sentences. This conversion allows established text mining techniques, such as classification methods to be applied to timeseries data. Half-hourly frequency surface dissolved oxygen data from 18 global lakes were used to create day-long segments of the original time series data. Three hundred sets of 1-d measurements were provided to a group of seven anonymous experts, experienced in manual filtering of oxygen data for metabolism modeling studies. The collective results were treated as expert panel decisions, and were used to rank the data by confidence level for use in metabolism calculations. While considerable variation occurred in the way the experts perceived the quality of the data, the model provides an objective and quantitative assessment method. The program output will assist the decision making process in determining whether data should be used for metabolism calculations. An R version of the program is available for download.

Ecosystem metabolism is an important and fundamental ecological concept. Many attempts have been made to numerically quantify its key components, productivity and respiration, for lake ecosystems across the world (Cole et al. 2000; Solomon et al. 2013). Ecosystem metabolism may be a proxy for trophic status and can be used to understand whether a lake is a source or sink of carbon (Hanson et al. 2003). As lake monitoring has become increasingly intensive and automated around the world (Weathers et al. 2013; Hamilton et al. 2015), the use of metabolism models to assess ecosystem functioning will likely grow.

Metabolism models in lakes typically assume that a change in free-water dissolved oxygen (DO) through time is driven primarily by the balance between photosynthesis (or primary production) and mineralization of organic carbon (often called “respiration” for simplicity), as well as equilibration of DO with the atmosphere (Staeher et al. 2010). When these three processes are dominant, diel DO patterns will be nearly sinusoidal, with increases during daylight due to primary production exceeding respiration and decreases at night due to respiration. However, additional processes, such as inflow and outflow to and from a lake, vertical and horizontal mixing, and advective movement of water mass can affect the balance of DO within specific lake strata (Antenucci et al. 2013; Rose et al. 2014) or between littoral and pelagic zones (Lauster et al. 2006; Van de Bogert et al. 2007; Batt and Carpenter 2012). When DO is measured using in situ sensors, these processes can impart patterns on the DO data that obscure the signal from biological processes and that, if left unaccounted, can

\*Correspondence: david.p.hamilton@griffith.edu.au

Additional Supporting Information may be found in the online version of this article.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

introduce noise and bias into the estimate of metabolism (Rose et al. 2014).

Generalizable approaches are needed for separating signals due to biological processes from those derived from physical processes to better quantify and debias lake metabolism estimates. For high frequency lake sensor observations, some attempts have been made to automate and standardize the methods of QA/QC (e.g., general QA/QC—Horsburgh et al. 2015) and calculation protocols (e.g., physical stability—Read et al. 2011; energy flux—Woolway et al. 2015; lake metabolism—Winslow et al. 2016). Experts have commonly removed data considered to be irrelevant noise or error, by visual assessment (e.g., Solomon et al. 2013), and in some cases have developed formalized approaches for evaluating uncertainty in metabolism predictions, as well as model parameters, and have identified the circumstances associated with those uncertainties (Cremona et al. 2014; Rose et al. 2014; Giling et al. 2017). While the aforementioned approaches have proven useful in evaluating metabolism predictions, they are subject to the overhead and constraints of coding parametric process-based models, and in some cases, the undocumented criteria of expert opinion. An alternative is to formalize the inclusion of expert knowledge on metabolism and use that knowledge, along with data-driven approaches, in efficient, flexible, and reproducible ways for data QA/QC.

Time series analysis, filtering, and data mining offer a set of solutions that may be particularly useful for evaluation of DO data intended for metabolism modeling (Niennattrakul et al. 2010; Rakthanmanon et al. 2011). Preparation for time series analysis should be comprised of three components operating either independently or simultaneously: QA/QC, data dimensionality reduction, and data representation/approximation. Increasing dimensionality (information), which is inherent in increased sampling frequency from sensors, decreases performance of similarity, or distance-based discovery algorithms (e.g., more difficult to build a robust model; Aggarwal et al. 2001; Zimek et al. 2012). This can be circumvented by removing some data or compressing the amount of information processed (Cannata et al. 2011) or by representing data in a simpler form (Keogh et al. 2001). Spectral analysis, such as Discrete Fourier Transformation (DFT) and Discrete Wavelet Transformation (DWT) are two examples that have been used in recent limnological contexts (e.g., Cengiz 2011; Kara et al. 2012; Cox et al. 2015). Techniques to accurately define “suitable data” have not been generalized but any methods needs to be robust and repeatable.

A promising technique that enables simplification of data while retaining key properties is Symbolic Aggregate approXimation (SAX; Lin et al. 2003). SAX has similarities to Piecewise Linear Approximation (PLA) and Piecewise Aggregate Approximation (PAA), which extract key information from complex time series data (Ralanamahatana et al. 2005). PLA and PAA divide time series into segments of equal or unequal length, and calculate segment trends or means for each segment. SAX

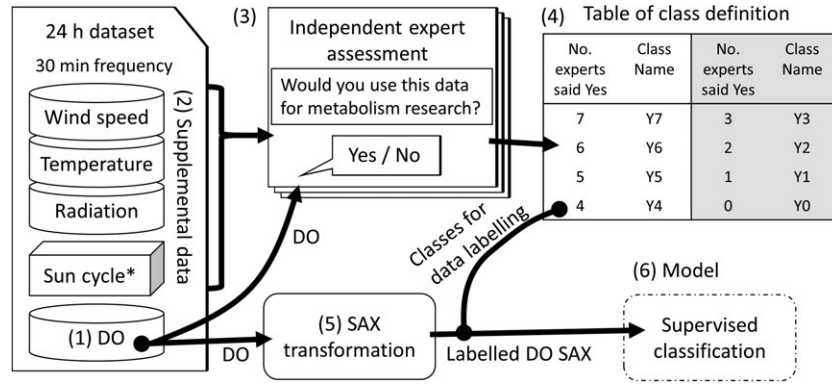
uses arithmetic mean values of even length segments (PAA), and further bins the segmented values into defined categories, creating a series of discrete letter sequences (words) from the original numeric time series (Lin et al. 2003, 2007). The SAX transformation enables the user to create a dictionary of time series subsequences similar to DNA sequences, making it possible to rapidly search for coherence in the time series vocabulary space. SAX analysis, due to its piecewise approach, is suitable for noisy and/or variable time series data common in environmental settings. SAX has been used in multiple disciplines such as vision based detection (e.g., Ma et al. 2016) and has recently been used in limnology to identify fluorescence signal patterns (Ruan et al. 2017).

The main objective of this study is to identify procedurally meaningful DO time series patterns from high frequency sensor data and provide a filter enabling identification and removal of complex data to improve the accuracy and consistency of lake metabolism calculations. The approach is designed to be reproducible and allow for automated classification of data quality that is consistent with expert opinion. To achieve this, the steps involved were: (1) generation of time series labels through expert evaluation, (2) transformation of time series data using the SAX method, and (3) supervised classification. We used a subsampled dataset from 18 lakes to generate and test the classification model.

## Methods

Eighteen lakes with suitable datasets (e.g., high-frequency preliminary QA/QCed surface DO, temperature profiles, and wind speed) for model training were selected from the Global Lake Ecological Observatory Network (GLEON) lakes. The majority of the data were reused from Solomon et al. (2013) (Table S1). The parent dataset contained 4852 d with dissolved oxygen data, ranging from 132 d to 434 d for individual lakes. To make labeling by experts feasible, random subsampling was used to obtain 300 d of data from the parent dataset, including 7–30 d (median 18 d) from individual lakes. For consistency, all time series data were downsampled to 30-min frequency.

Seven scientists at a conference were approached for their expertise in lake metabolism studies, i.e., experience with screening these datasets. The 300 d of subsampled data were provided to the experts as time series of DO over each day. Also included were supplementary figures comprising time series of water column temperature profile, wind speed, and photosynthetically active radiation (PAR), as well as the timing of sunrise and sunset, as these data could be used to further inform the experts about the quality of the data and the relevant processes. The group members were asked to evaluate which specific days of DO time series data were suitable for lake metabolism analysis, based on their experience and inspection of the visualized dataset. Three questions were asked of the experts in relation to each dataset: (Q1) “Would



**Fig. 1.** The workflow for generation of the classification model. Three hundred days of dissolved oxygen concentration (DO) at 30 min frequency were provided (1) to seven independent experts, along with supplementary data (2). Experts labeled the data (3), which was then collated and allocated according to classes (Y7–Y0) representing the number of experts that said “Yes” to the data being useful (4; answers “maybe yes” and “maybe no” were aggregated to Yes and No, respectively). The identical 300 d of DO time series data were also transformed (5) by Symbolic Aggregate approXimation (SAX), and (6) a classification model was created using (5) to reproduce the labels (4). Sun cycle includes sunrise and sunset timing.

you use this DO data for metabolism studies?”, (Q2) “Did biological processes dominate the metabolism signal represented in DO?”, and (Q3) “Other than DO, what data influenced your Q1 decision?”. Four choices were provided as options to Q1, namely [Yes], [Maybe Yes], [Maybe No], and [No]. The responses of the experts were aggregated and turned into labels for each day, based on eight possible classes: Y0, Y1, Y2, Y3, Y4, Y5, Y6, and Y7. For example, data were labeled as Y7 (best class) if all seven scientists selected either options [Yes] or [Maybe Yes], and as Y0 (worst class) if all scientists selected [No] or [Maybe No]. This method was used to provide an independent quantitative expert evaluation of the level of confidence in the quality of the data. The survey results were analyzed according to: (Q1) frequency of expert agreement, (Q2) whether usability of DO data was related to the dominance of biological processes in the DO signal, and (Q3) whether experts indicated additional data would have helped to refine Q1.

Labels Y0–Y7 from the expert panel assessment were used to build classification models after SAX transformation of each day of data. A diagram of this process is shown in Fig. 1. The classification models used R libraries rWeka (ver. 0.4-34; Witten et al. 2016; Hornik et al. 2009), RWeKajars (ver. 3.9.1-3; Hornik 2018), RJava (Simon Urbanek 2017) and shiny (ver. 1.0.3; Chang et al. 2017) on R (ver. 3.4.1). R was selected as the main framework since it is widely used in ecology and data mining disciplines and is open-source software. WEKA (Waikato Environment for Knowledge Analysis) is specialized data mining and machine learning software (Hall et al. 2009). Both rWeka and RWeKajars are APIs (application program interface) in the R language platform that enable use of a variety of data mining resources through WEKA toolsets. Shiny is a library allowing the creation of a user-friendly front-end for the models.

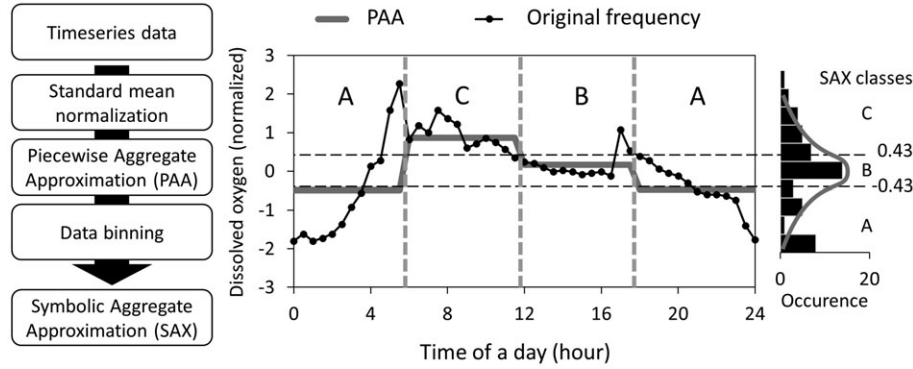
To prepare data for the supervised classification, we followed the protocol described by Lin et al. (2007) for the SAX transformation. The SAX transformation combines two time

series transformation methods that reduce the dimensionality of the data: piecewise averaging and data binning (Fig. 2). The piecewise averaging method, also known as Piecewise Aggregate Approximation (PAA), segments the original time series data (measured at a 30-min resolution) into  $n$  equal time periods for which an average value is derived. For example, PAA applied to a 24-h DO dataset with  $n = 4$  will contain an average value for each of the four 6-h time segments. Similarly, data binning (into  $m$  bins) was used to segment DO values. For example, with  $m = 2$  data bins, DO values can be defined as being  $\geq$  or  $<$  a specified breakpoint value (Table sB). The binned data therefore holds ordinal information rather than nominal or numeric values. Each bin is represented by a letter in the processed version of the time series so that the original numeric times series becomes an alphabetic string.

The SAX transformation was carried out after normalizing the original DO daily timeseries using a standard mean transformation:

$$\text{DO}_{\text{norm}} = (\text{DO} - \mu) / \sigma \quad (1)$$

(1) where  $\text{DO}_{\text{norm}}$  is the normalized DO time series,  $\mu$  is the arithmetic mean of DO and  $\sigma$  is the standard deviation of DO for the day. Breakpoints were identified by splitting the  $\text{DO}_{\text{norm}}$  values into equal percentile probabilities assuming a standard normal distribution (see Table sB). Once the  $m-1$  breakpoints were identified using PAA, and thus a mapping from  $\text{DO}_{\text{norm}}$  values to letters of an alphabet with  $m$  letters established,  $\text{DO}_{\text{norm}}$  was averaged for each of the  $n$  time periods and turned into alphabetic representation by looking up the appropriate bin in the list of  $m$  bins. The lowest numerical values of  $\text{DO}_{\text{norm}}$  were given the letter “a.” Assuming, for example,  $m = 3$ , the largest numerical values would be given the letter “c.” We express a SAX transformation with  $n$  PAA segments (corresponding to the size of the “words” that will represent each time series) and  $m$  bins (the size of the



**Fig. 2.** Schematic of the SAX transformation. The graph (middle) shows an example of normalized dissolved oxygen (DO\_norm) data at 30 min intervals (black line with dots), its PAA results at 6 h intervals (thick vertical gray dashed lines) and SAX letters according to the breakpoints given in Table sB (dashed lines; 0.43 and -0.43). In this example, the SAX word length ( $n$ ) is 4 and there are 3 letters ( $m$ ) corresponding to the two breakpoints. The right histogram shows the distribution of the data with the gray line representing an idealized normal distribution. The SAX transformation processes are shown on the left-hand side. In this case, the data consists of the following SAX letter combinations and counts (shown by numbers): [A (2), B (1), C (1), AC (1), BA (1), CB (1), ACB (1), CBA (1), ACBA (1)]. The classification model uses the set of subsequence counts as input variables for logistic regression.

alphabet) as  $SAX(n, m)$ . We deployed 25 SAX parameter sets ( $n = 2-6$ ;  $m = 2-6$ ) to examine performance of SAX against expert opinion. An R algorithm by Ruan et al. (2017), which uses the classic SAX technique, was used.

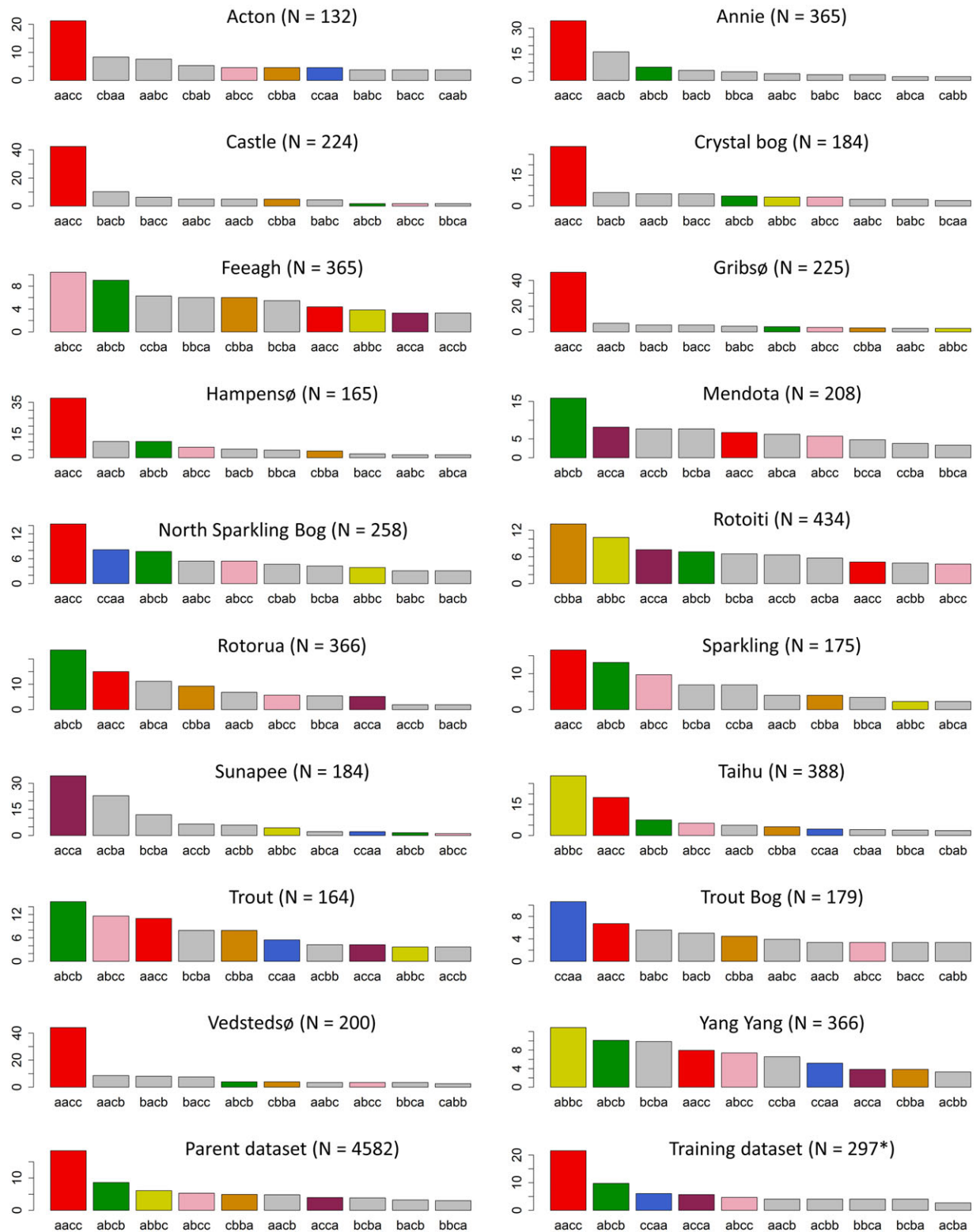
In our study, we use the SAX-transformed time series to formulate a classification problem by associating each transformed series with one of the eight labels (Y0–Y7) generated from expert input. More specifically, we created an ordinal classification problem because the eight labels exhibit a natural order. Standard supervised learning algorithms cannot exploit this ordering information without converting the classes into numeric values. To overcome this issue, our model creates the following seven two-class problems: [Y0 | Y1 – Y7], [Y0 – Y1 | Y2 – Y7], [Y0 – Y2 | Y3 – Y7], [Y0 – Y3 | Y4 – Y7], [Y0 – Y4 | Y5 – Y7], [Y0 – Y5 | Y6 – Y7], and [Y0 – Y6 | Y7] where the threshold “|” separates the first and second binary class, i.e., unsuitable and suitable data respectively. For brevity, we use the notation Y0-1 to refer to the two-class problem [Y0 | Y1 – Y7], and so on for other classes. Based on this model setting, class probability estimates from the seven two-class models, one for each threshold, were combined to obtain multi-class probability estimates for all eight categories for each test sequence, assigning the sequence to the class with maximum probability. The method proposed by Frank and Hall (2001), in conjunction with the smoothing method from Schapire et al. (2002), was used to combine the two-class probability estimates into multi-class probability estimates. This process was implemented in the OrdinalClassClassifier procedure that is available in R via RWeka. To compare the sensitivity of SAX parameters to the model performance, we examined the model performance using the seven two-class problems.

Logistic regression, the classification technique we apply to our data, requires numeric input rather than strings of letters. We established the numeric features by computing subsequence frequencies for each sequence of letters to be classified. More specifically, for a  $SAX(n, m)$  model, which generates

strings of length  $n$  consisting of  $m$  letters, we count how often each of the  $\sum_{i=1}^n m^i$  theoretically possible subsequences occurs in the sequence to be classified (as we only considered subsequences consisting of consecutive letters). The set of subsequence counts are used as the predictor variables in the logistic regression model.

Due to the available SAX parameter combinations, twenty-five candidate models were generated and tested for their performance. The model performance was evaluated in the form of the binary classes (suitable and unsuitable) for each of the seven two-class problems discussed above. To measure performance, we used Area Under the ROC Curve (AUC), and Matthews Correlation Coefficient (MCC). We also considered a confusion matrix for the classification problem to obtain additional insight. A confusion matrix is a frequency distribution table of the test data instances, illustrating how instances of class X are assigned to class Y by the classification model. A confusion matrix for a two-class problem shows the following frequencies: TP (True Positives), TN (True Negatives), FP (False Positives), and FN (False Negatives).

A receiver operating characteristics (ROC) curve shows the true positives rate ( $TPR = TP / (TP + FN)$ ) and false positives rate ( $FPR = FP / (FP + TN)$ ) in two-dimensional space (Bradley 1997; Witten et al. 2016). Each TPR/FPR point in this space is obtained by applying a different classification threshold on the class probability estimates obtained from the classification model. To summarize the information in the curve, the area under the curve (AUC) is used as a performance measure. It can be shown that AUC corresponds to the estimated probability that a randomly chosen positive test instance is ranked above a randomly chosen negative test instance when the classifier's class probability estimates for the positive class are used to rank the test instances. AUC is less sensitive to the relative frequency of the two classes (positive and negative) than simple TPR or FPR measures, allowing direct comparison



**Fig. 3.** The 10 most frequently recurring sequences of daily DO SAX letters from 18 lakes as well as parent and training (subsampling 300 d) datasets are shown in proportion to the entire data used (Y axis: frequency of occurrence). For this, SAX transformation was parameterized with SAX(4,3); three letters (a, b, c) and four segments a day. Theoretically there are  $3^4 = 81$  possible sequences. The seven sequences that occurred most frequently across the set are highlighted with colors to aid intuitive recognition of their frequency of detection (aacc-red; abcc-pink; abcb-green; cbba-orange; acca-violet; abbc-yellow; ccaa-blue). Parent (all lake) and training datasets are also shown. Two sequences abbc-yellow and cbba-orange that did not show up in the top 10 training data have instances of seven and eight, respectively appearing in the training data.

**Table 1.** (A) Percentages (%) of full day DO SAX( $n,m$ ) sequences that appeared in the parent dataset ( $N = 4582$ ) in comparison to all the possible combination of letters ( $m^n$ ) in various number of word size ( $n$ ) and alphabet ( $m$ ) settings. (B) Percentages of full day DO SAX( $n,m$ ) unique sequences that appeared in the training dataset in comparison to parent dataset patterns in various number of word size ( $n$ ) and alphabet ( $m$ ) settings. (C) Percentages of parent data incidents (i.e., number of days of  $N = 4582$ ) covered by training dataset in terms of SAX sequence.

|   |   | Alphabet size |      |      |      |      |
|---|---|---------------|------|------|------|------|
|   |   | 2             | 3    | 4    | 5    | 6    |
| <b>A: Parent dataset coverage (sequence)</b>    |   |               |      |      |      |      |
| Word size                                       | 2 | 75.0          | 77.8 | 50.0 | 52.0 | 33.3 |
|   | 3 | 75.0          | 70.4 | 39.1 | 43.2 | 27.3 |
|   | 4 | 87.5          | 66.7 | 47.7 | 36.0 | 26.2 |
|   | 5 | 93.8          | 63.8 | 38.4 | 22.2 | 13.1 |
|   | 6 | 92.2          | 48.8 | 21.0 | 9.0  | 4.1  |
| <b>B: Training dataset coverage (sequence)</b>  |   |               |      |      |      |      |
| Word size                                       | 2 | 66.7          | 71.4 | 62.5 | 61.5 | 50.0 |
|   | 3 | 100.0         | 94.7 | 92.0 | 72.2 | 78.0 |
|   | 4 | 100.0         | 72.2 | 57.4 | 45.3 | 38.2 |
|   | 5 | 76.7          | 49.7 | 31.8 | 22.8 | 19.0 |
|   | 6 | 61.0          | 28.7 | 18.4 | 15.3 | 12.5 |
| <b>C: Training dataset coverage (incidents)</b> |   |               |      |      |      |      |
| Word size                                       | 2 | 100.0         | 99.9 | 99.9 | 99.9 | 99.8 |
|   | 3 | 100.0         | 99.8 | 99.6 | 96.7 | 96.3 |
|   | 4 | 100.0         | 96.8 | 90.2 | 82.2 | 76.7 |
|   | 5 | 97.7          | 87.9 | 75.4 | 58.5 | 52.6 |
|   | 6 | 96.8          | 78.7 | 60.2 | 46.8 | 29.1 |

across different threshold settings. Model performance is considered to be perfect if AUC = 1, and random if AUC = 0.5.

Matthews correlation coefficient (MCC, proposed by Matthews 1975) is an alternative accuracy classification that is not affected by imbalanced class distributions. MCC is a discrete version of the Pearson correlation coefficient, varying from 1 (perfect fit) to 0 (no fit). Negative values are also possible if “anti-learning” has occurred. MCC is calculated as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (2)$$

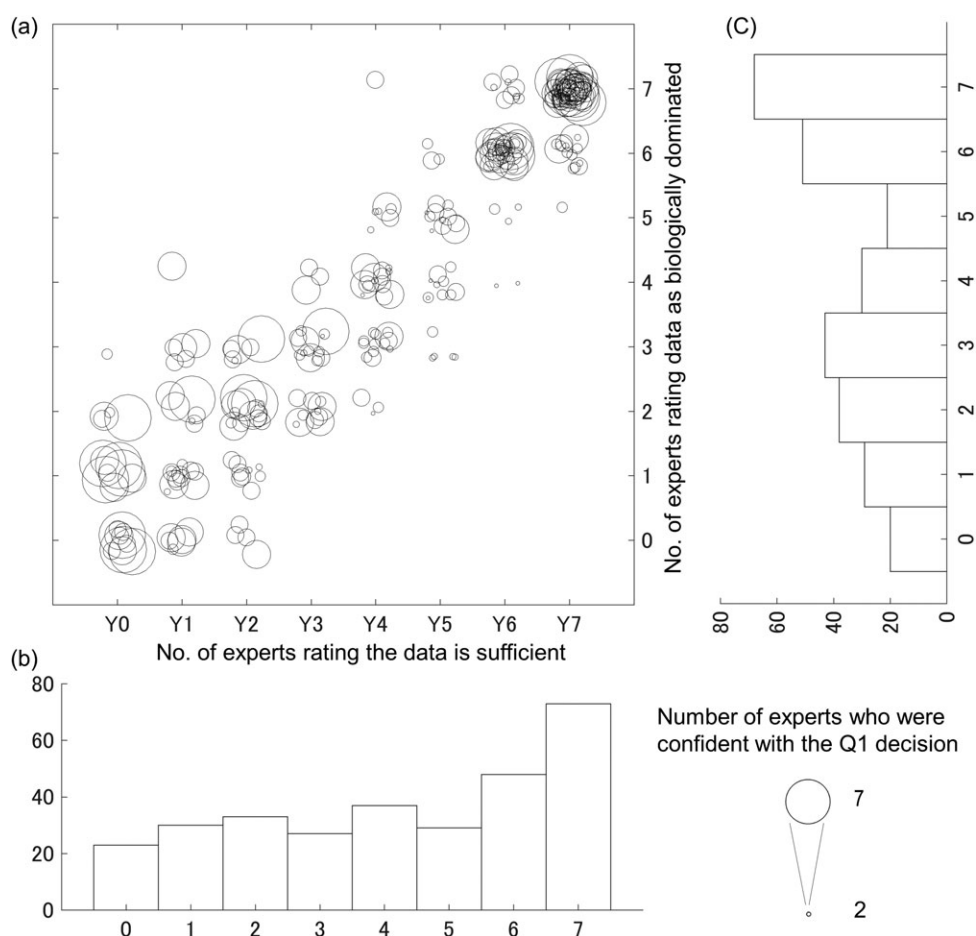
Both AUC and MCC information were used to determine appropriate models. Evaluation of AUC and MCC was carried out in a 10-fold cross-validation process to estimate performance of the full data model; i.e., they were estimated by the average of the 10 results obtained from a rotated 10% data split validation (Kohavi 1995). These model evaluations were examined in the two-class models, while the full confusion matrix provided insights into the combined multi-class model.

## Results

### Data exploration and subsampling

The 10 most frequently recurring daily DO SAX sequences of the parent dataset (4582 d of DO data) are shown in Fig. 3.

The results are from the SAX(4,3) transformation (“candidate models” section explores different SAX transformation results), i.e., with 6-h resolution ( $n = 4$ ) and two thresholds ( $m = 3$ ). Recurrent patterns occur across several of the lakes and most of these patterns start with the letter “a” (i.e., the bin with the lowest normalized DO). The sequence “aacc” [i.e., DO is low in the first half of the day (0–12 h) and high in the second half of the day (12–24 h)] is the most frequently occurring pattern in nine lakes and “abcb” [i.e., DO rises through the first three quarters of the day (0–18 h), and then decreases in the fourth quarter (18–24 h)] is the most frequent pattern in three lakes. The letters are not randomly distributed, suggesting the feasibility of categorization of daily DO observations based on letter sequences alone. While the SAX(4,3) transformation theoretically results in  $3^4 = 81$  possible full day sequences, the parent dataset includes only 54 (66.7%) of these patterns (Table 1A). Considering the substantial size of the parent data used, the parent data patterns in small SAX parameters are thought to include all idealized DO curves driven by biological activities, and therefore those theoretical patterns that did not appear in the parent datasets are primarily “noisy.” This coverage decreases as the number of possible SAX strings increases. The lowest coverage is found in SAX(6,6), where 4% of the available sequences appeared in the parent data. An exception occurs for  $m = 2$ , where the parent dataset



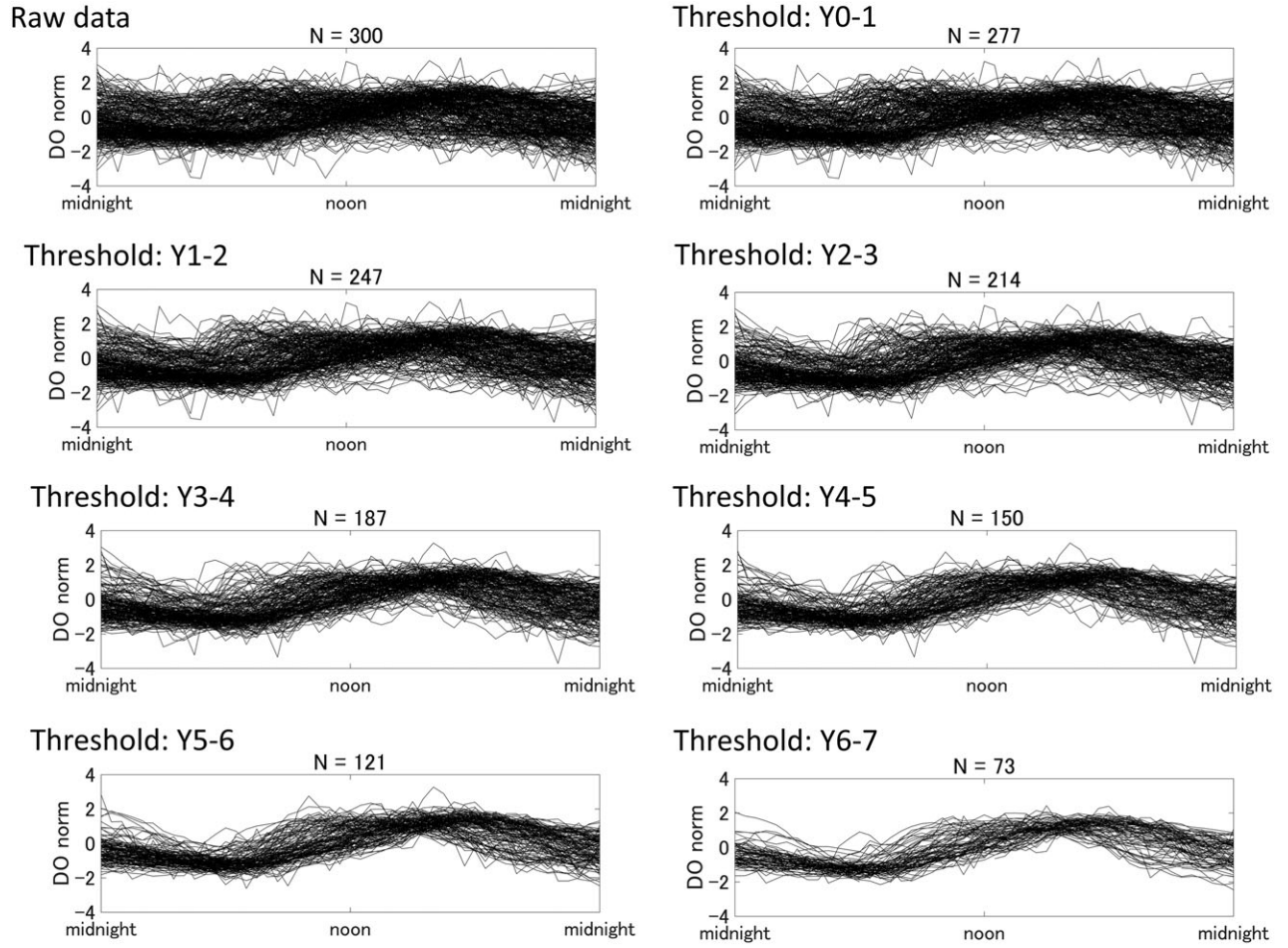
**Fig. 4.** (a) Scatter diagram: number of experts indicating that daily data is biologically dominated vs. data adequacy (Y0–Y7) based on number of experts indicating “Yes.” Circles are plotted with a small degree of randomness (0.25 jitter) to reduce visual data overlap of the discrete values, and size of the circles reflects the number of experts who were confident with their individual decision (Pearson’s correlation coefficient: 0.87;  $p < 0.01$ ). Histograms complement the scatterplot to indicate frequency distribution of experts indicating that DO data were adequate (b) and that DO data were biologically dominated (c).

coverage generally increases when  $n$  increases. It is noteworthy that the differences in coverage are primarily determined by alphabet size rather than word length.

The coverage of sequences observed in the subsampled parent dataset (300 d) is summarized in Table 1B. For SAX (4,3), 39 (72%) of the 54 sequences identified in the parent dataset are present. A higher proportion of parent data patterns are covered in the subsampled data when both SAX parameters are small. Both alphabet size and word length similarly affect the training (subsampled 300 d) data coverage of the parent data. Table 1C shows the proportion of parent data full-day SAX sequences represented in the training data. For the SAX(4,3) setting, over 95% of parent data sequences are represented, leaving 149 instances not represented in the training data. Similarly, the majority of parent data sequences are included for most of the SAX parameter settings, while two SAX parameter settings (SAX(6,5) = 47%; SAX(6,6) = 29%) fail to represent more than one-half of the instances.

### Survey results

The seven experts rated 34–80% (average 60%) of the 300 daily training data as suitable for lake metabolism analysis. For the threshold separation Y3-4 ([Y0 – Y3 | Y4 – Y7]), an average of 62% of the training data was labeled as “suitable” (Fig. 4). The highest number of data instances was recorded in Y7 ( $n = 73$ ), with 32 instances on average for the other classes (min = 23, max = 48, Fig. 4). Figure 5 illustrates “suitable” 30-min DO data according to the expert panel results and assigned thresholds. The available “suitable data” reduces as the threshold level increases, but it is evident that noise in the data are filtered out through the expert panel evaluation process. The experts chose options [Yes] and [No] without “maybe” 85% of the time, while classes Y3–Y5 contained more “maybe” responses. Survey results for Q1 (Would you use this DO data for metabolism studies?) and Q2 (Did biological processes dominate the metabolism signal represented in DO?) were strongly positively correlated. The survey results for Q3 (Other than DO, what data influenced your Q1 decision?)



**Fig. 5.** The “good data” consisting of 30 min interval timeseries over 1 d and classified according to the expert panel decision. Seven thresholds are shown.  $N$  represents number of days that were classified as having “good data.”

indicated that 29.6% of the time, the majority of the experts used one or more supplementary data sources for their assessment, but the type and number of supplementary data varied. The number of times the panel requested additional data was 20 for PAR, 37 for wind speed, 1 for diel solar radiation, 0 for surface temperature, 48 for temperature profile, and 0 for other information. On 13 occasions, the panel cited two additional sources of supplementary data as being required to make their decision on Q1.

### Candidate models

Parameter selection and threshold analysis were attempted through a model selection process. Twenty-five models were created based on different SAX parameter combinations (SAX( $n, m$ )), i.e.,  $n = 2-6$  corresponding to 12–4 h intervals, respectively, and  $m = 2-6$  corresponding to one to five threshold values to separate DO data. Tables 2–4 show 10-fold cross validation results of the model performance using MCC and AUC metrics. We examined all possible combination of SAX parameters as well as thresholds, while shown here are what we

consider useful information. Table 2 gives the results of various models when the threshold was set to Y5-6. Table 3 provides the results with the SAX alphabet number fixed to 3, i.e., SAX( $n, 3$ ), and Table 4 indicates the results with SAX word length fixed to 4, i.e., SAX(4,  $m$ ). For Y5-6 ([Y0-Y5 | Y6-Y7]), only SAX(4,3) appeared among the top five results based on both MCC and AUC analyses. For the different threshold settings using SAX( $n, 3$ ),  $n = 4$  performed better in both MCC and AUC analyses. For the models with SAX(4, $m$ ),  $m = 3$  results ranked in top five performance in both MCC and AUC analysis.

An extended confusion matrix using SAX(4,3) is shown in Fig. 6. For example, for the threshold Y3-4 ([Y0-Y3 | Y4-Y7]), 174 instances were correctly classified as “suitable data” (TP) and 13 instances were wrongly classified as “unsuitable data” (FN). This means that of the 187 instances of “suitable data” (for the Y3-4 threshold), the model correctly labeled 93% instances. Conversely, TN = 85 and FP = 28, which means 75% of the “unsuitable data” was correctly classified as unsuitable.



**Table 2.** Ten-fold cross validated model performances in terms of Mathews correlation coefficient (MCC, top) and area under the receiver operating characteristic curve (AUC, bottom) with various SAX word sizes and size of SAX alphabet, where threshold was fixed to Y5-6 ([Y0-Y5 | Y6-Y7]). Numbers in bold represent the top five results in the table.

| MCC       |   | Alphabet size |             |             |             |             |
|-----------|---|---------------|-------------|-------------|-------------|-------------|
|           |   | 2             | 3           | 4           | 5           | 6           |
| Word size | 2 | 0.53          | 0.45        | 0.35        | 0.44        | 0.40        |
|           | 3 | 0.28          | 0.41        | 0.44        | 0.41        | 0.34        |
|           | 4 | 0.47          | <b>0.63</b> | 0.56        | 0.47        | 0.55        |
|           | 5 | 0.53          | <b>0.59</b> | 0.51        | 0.51        | <b>0.64</b> |
|           | 6 | <b>0.64</b>   | 0.49        | 0.52        | <b>0.58</b> | 0.53        |
| AUC       |   | Alphabet size |             |             |             |             |
|           |   | 2             | 3           | 4           | 5           | 6           |
| Word size | 2 | 0.72          | 0.73        | 0.76        | 0.74        | 0.75        |
|           | 3 | 0.72          | 0.73        | 0.77        | 0.75        | 0.72        |
|           | 4 | 0.79          | <b>0.88</b> | <b>0.87</b> | 0.80        | 0.82        |
|           | 5 | 0.81          | 0.84        | 0.81        | 0.80        | <b>0.88</b> |
|           | 6 | 0.83          | <b>0.85</b> | <b>0.85</b> | 0.82        | 0.84        |

Figure 7 shows six normalized DO time series of instances with extreme errors (i.e., expert labels Y0 – Y2 were classified as Y7). The fact that the classifier mis-classifies these SAX sequences (aacc, abcb, bcca) implies that they appeared repeatedly in the training dataset and their corresponding DO time series were frequently identified by the expert panel as Y7. Inspection of the plots in Fig. 7 shows that the likely reasons for the mis-classifications are: (1) appearance of repeated values over a part of the day, (2) low variations of DO values, and (3) obvious increase in DO before sunrise. Figure 8 illustrates DO data at 30-min intervals for days when data are classified as “suitable” according to the SAX(4,3) model. The SAX(4,3) model generally overestimated the amount of “suitable data” in each threshold compared with the expert panel labels (Fig. 5). The number of instances of “suitable data” classified into the higher threshold levels was greater than the expert panel decisions in favor of those thresholds (i.e., errors for Y5-6 and Y6-7 were 31 and 45, respectively), but the number of classifications is similar for the lower thresholds (mean error for Y0-1 to Y4-5 was 14).

## Discussion

Humans have a great capacity for detecting visual patterns (e.g., Cox et al. 1997), and our approach to evaluating the suitability of DO data for metabolisms models exploits that capacity. As previous work has shown, DO time series are often messy and have complex patterns, and teasing-apart the underlying causes of noise and bias in metabolism estimates made from DO signals is extremely challenging (Cremona et al. 2014; Rose et al. 2014; Giling et al. 2017). Our method provides an alternative to the more classical approach of parametric analysis by basing the classification of DO signals on expert knowledge,

**Table 3.** Ten-fold cross validated model performances in terms of Mathews correlation coefficient (MCC, top) and area under the receiver operating characteristic curve (AUC, bottom) with various SAX word sizes and threshold settings, where size of SAX alphabet was fixed to 3. Numbers in bold represent the top five results in the table.

| MCC       |   | Threshold |      |      |             |             |             |      |
|-----------|---|-----------|------|------|-------------|-------------|-------------|------|
|           |   | Y0-1      | Y1-2 | Y2-3 | Y3-4        | Y4-5        | Y5-6        | Y6-7 |
| Word size | 2 | 0.13      | 0.34 | 0.49 | 0.51        | 0.43        | 0.45        | 0.00 |
|           | 3 | 0.30      | 0.51 | 0.46 | 0.43        | 0.50        | 0.41        | 0.00 |
|           | 4 | 0.29      | 0.50 | 0.52 | <b>0.61</b> | <b>0.63</b> | <b>0.63</b> | 0.36 |
|           | 5 | 0.23      | 0.48 | 0.46 | 0.58        | <b>0.59</b> | <b>0.59</b> | 0.26 |
|           | 6 | 0.34      | 0.42 | 0.47 | 0.59        | 0.50        | 0.49        | 0.27 |
| AUC       |   | Threshold |      |      |             |             |             |      |
|           |   | Y0-1      | Y1-2 | Y2-3 | Y3-4        | Y4-5        | Y5-6        | Y6-7 |
| Word size | 2 | 0.65      | 0.71 | 0.74 | 0.77        | 0.74        | 0.73        | 0.71 |
|           | 3 | 0.74      | 0.81 | 0.78 | 0.75        | 0.72        | 0.73        | 0.67 |
|           | 4 | 0.71      | 0.71 | 0.82 | <b>0.84</b> | <b>0.86</b> | <b>0.88</b> | 0.77 |
|           | 5 | 0.71      | 0.77 | 0.77 | 0.83        | 0.81        | <b>0.84</b> | 0.76 |
|           | 6 | 0.65      | 0.62 | 0.71 | 0.76        | 0.81        | <b>0.85</b> | 0.77 |

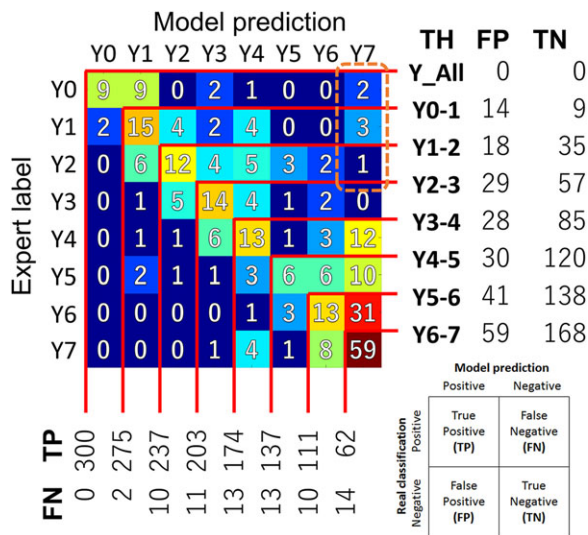
as well as the patterns inherent in the DO data. The aforementioned approaches focus on the suite of processes, whereas our approach focuses on the suite of patterns.

## A framework for labeling data

Our primary goal was to design a framework to provide simple labeling of suitable and unsuitable (i.e., suitable and unsuitable for free surface metabolism models) segments in high frequency autonomous DO data. Black-and-white expert panel decisions were not made, as the experts had different interpretations and expectations about the data provided to them. Consequently, our model provides a semi-qualitative, but informative judgment about: “how many experts would support the quality of data,” by reproducing labels Y0–Y7. A user is then given the freedom to choose a threshold decision level based on their expectation of confidence in the data quality and the number of suitable data available for analysis. While this leaves a degree of variation in the data products, scientists may have different expectations for cleaning data. For a metabolism model study, for example, if a higher threshold (such as Y6-7) is used, the classification model may output data that has mostly idealized shapes of DO over a diurnal cycle (i.e., alternation of dominance by production and respiration). Such a case may be expected to occur where changes from transport and mixing are of lesser significance than biological processes. The selection of high levels of confidence might, however, restrict the number and frequency of data available for use by the metabolism model, and may well disregard specific features that occur in reality. Conversely, if a lower threshold such as Y3-4 is used, more data will be available, but the user will need to take a cautious approach to the

**Table 4.** Ten-fold cross validated model performances in terms of Mathews correlation coefficient (MCC, top) and area under the receiver operating characteristic curve (AUC, bottom) with various size of SAX alphabet and threshold settings, where SAX word size was fixed to 4. Numbers in bold represent the top five results in the table.

| MCC                 |   | Threshold |      |      |             |             |             |      |
|---------------------|---|-----------|------|------|-------------|-------------|-------------|------|
|                     |   | Y0-1      | Y1-2 | Y2-3 | Y3-4        | Y4-5        | Y5-6        | Y6-7 |
| Number of alphabets | 2 | 0.24      | 0.54 | 0.57 | <b>0.59</b> | 0.46        | 0.47        | 0.07 |
|                     | 3 | 0.29      | 0.50 | 0.52 | <b>0.61</b> | <b>0.63</b> | <b>0.63</b> | 0.36 |
|                     | 4 | 0.17      | 0.48 | 0.49 | 0.58        | <b>0.67</b> | 0.56        | 0.30 |
|                     | 5 | 0.24      | 0.37 | 0.42 | 0.52        | 0.49        | 0.47        | 0.36 |
|                     | 6 | 0.25      | 0.36 | 0.46 | 0.33        | 0.53        | 0.55        | 0.26 |
| AUC                 |   | Threshold |      |      |             |             |             |      |
|                     |   | Y0-1      | Y1-2 | Y2-3 | Y3-4        | Y4-5        | Y5-6        | Y6-7 |
| Number of alphabets | 2 | 0.62      | 0.72 | 0.78 | 0.82        | 0.77        | 0.79        | 0.74 |
|                     | 3 | 0.71      | 0.71 | 0.82 | <b>0.84</b> | <b>0.86</b> | <b>0.88</b> | 0.77 |
|                     | 4 | 0.63      | 0.71 | 0.82 | <b>0.84</b> | 0.83        | <b>0.87</b> | 0.79 |
|                     | 5 | 0.69      | 0.63 | 0.75 | 0.82        | 0.80        | 0.80        | 0.78 |
|                     | 6 | 0.72      | 0.74 | 0.76 | 0.75        | 0.80        | 0.82        | 0.74 |

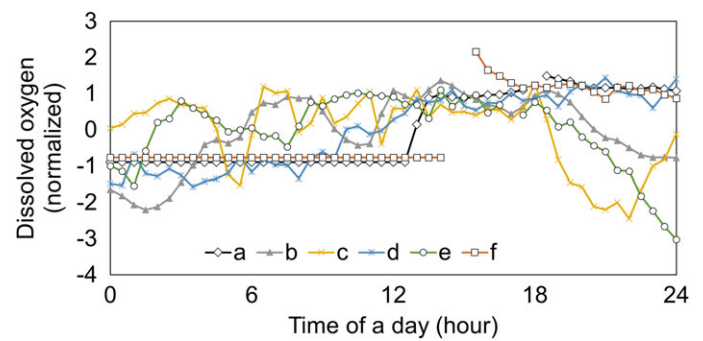


**Fig. 6.** All training data model results for eight classes Y0–Y7 in relation to the extent of expert agreement, where Y7 (y-axis) corresponds to the full consensus on the use of the data. Red lines illustrate the binary class threshold settings, and for each threshold, True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) were calculated. The schematic figure at the bottom right shows the basic structure of a confusion matrix for a two-class problem. TH stands for class threshold, and extreme errors (orange dashed box) are explored in Fig. 7. The color was added to provide visual realization of the number.

interpretation of results since by definition, reducing the Y value reflects reduced levels of expert panel confidence in the data quality.

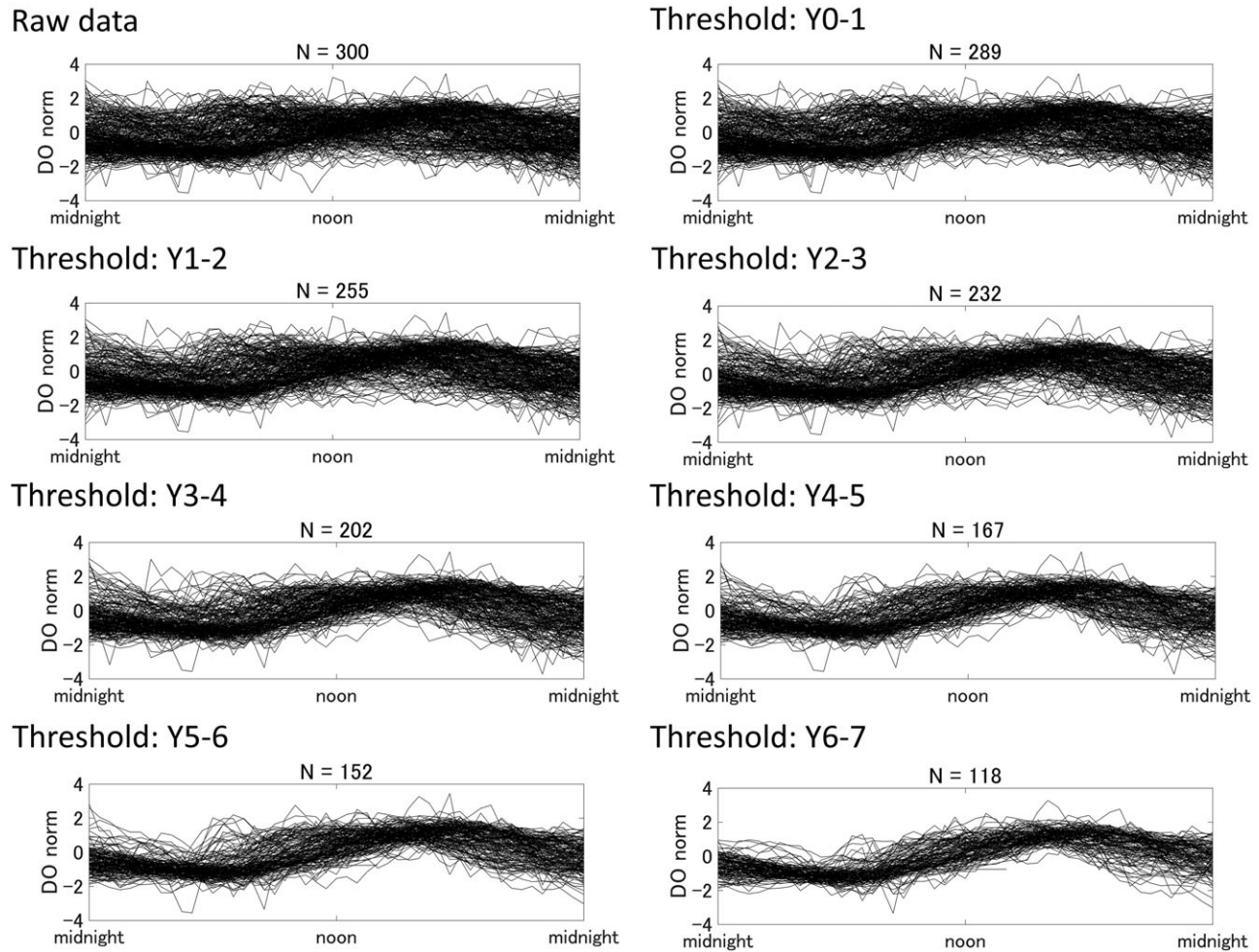
#### SAX as transformation for data QA/QC and analysis

With increasing use of autonomous in situ sensors, and the resulting large volume and high complexity of observations, it



**Fig. 7.** Six timeseries of normalized DO indicated as extreme errors in Fig. 6. SAX(4,3) for each timeseries were: (a) aacc; (b) abcb; (c) bcca; (d) aacc; (e) bcca; (f) aacc. Variations of DO in  $\text{mg L}^{-1}$  (max - min) for each series were (a) 0.28; (b) 1.13; (c) 0.38; (d) 0.07; (e) 0.90; (f) 2.03, and standard deviations were (a) 0.12; (b) 0.32; (c) 0.10; (d) 0.02; (e) 0.22; (f) 0.69. Suspected causes of errors include: repeated values (a, f), increase of DO before sunrise (b, e), and low variation of data (a, c, d).

is increasingly difficult to manage, archive and analyze data. Ecologists would therefore benefit from embracing approaches that meld simple models with machine learning. The conventional QA/QC approach for evaluating data from autonomous sensor networks is no longer practical due to the rapid expansion of sensors, networks and “big data” generally (Campbell et al. 2013). Common QA/QC tasks typically focus on network or sensor malfunctions, such as missing values, sensor drift, or inconsistency. Observations of DO can be susceptible to these types of malfunctions, but further filtering is necessary for more complex tasks, for example, assessment of lake metabolism. SAX is a simple transformation of time series data. The transformed data also provides a different way to think about environmental data as a sequence of words, and the unique



**Fig. 8.** Dissolved oxygen data classified as “good” for 30-min interval timeseries over 1 d according to the SAX(4,3) model results and with seven thresholds.  $N$  is the number of data classified as “good data.”

approach opens up opportunities for additional analytical tools, such as a text sequence mining approach to analyze (dis)similarity of sequence patterns.

DO signal variations are known in both intra-lake data (driven mostly by seasonality) and inter-lake data (driven mostly by latitude, trophic status, and geology) (e.g., Richardson et al. 2017). These variations might be evident as different magnitudes of variation and frequency of peaks and troughs in the data, and could also be influenced by differences in sunrise/sunset timing, or complex balances of productivity, respiration, transportation, diffusion, and surface gaseous fluxes. The latter information (i.e., magnitude and rate of a change) was explicitly not used to filter out the data in our classification model, as this would potentially give bias to the metabolism model. Many other variables may also require complex QA/QC processes, as well as filtering, to make sense of the data. For example, phycocyanin sensors require consideration of factors that affect the assumed linearity between cyanobacteria biomass and phycocyanin, such as the proportion of

colonial or filamentous populations, temperature or species-specific signals (Chang et al. 2012). In other words, relationships between environmental sensor readings and the data of interest may require specific knowledge or sensor conversions to assess and extract information relevant to the variable of interest (Kara et al. 2012).

Classification of time series data is challenging as the data usually exhibit high dimensionality and are inherently noisy (Keogh and Kasetty 2002; Hanson et al. 2008). To overcome this, a classification model should only be provided with appropriate information extracted from the data. For our case, the essence of the data is the shape of the time series. The variations of DO peaks and troughs and the timing of these requires a robust analytical procedure. The SAX transformation is simple in concept but has proven useful in many applications (Lin et al. 2007). In essence, SAX removes quantitative uncertainties and only preserves the general shape of time series data. Symbols, as a result of data normalization and binning in SAX, are equiprobable. In other words, the probability

of occurrence of each symbol is likely to be equal, on the assumption that the values in the time series are normally distributed. This provides good coverage when using SAX to detect shapes and trends in sequences by applying string sequence classification methods (Ralanamahatana et al. 2005). SAX symbols are robust and clean due to the segmented approximation process (PAA). As a result, most of the variability in the observed data can be represented semi-quantitatively. PAA also reduces the computational memory and run time used for classification, and thus allows for comparison of multiple models in a computationally efficient way. In the most extreme case in this study, data were reduced in number from 1440 samples per day to 4 [in case of SAX(4,3)].

### Generalizability of the SAX and expert opinion approaches

The SAX transformation of DO for 18 lakes resulted in discovery of relatively consistent diurnal DO sequences across most lakes. We emphasize that our evaluation is not of the lakes, per se, but of the collection of DO patterns likely to be encountered at the daily scale that have relevance to metabolism. To be clear, we are not evaluating the mean or variance of DO, but rather the specific patterns. Given a dimensionality of SAX(4,3) which was found to be most effective at reproducing expert classification, there are 81 possible patterns. Only about one half of the patterns were found in the data, and of these, 7–10 patterns accounted for most of the occurrences (Fig. 3). Put another way, a relative few patterns account for most of the patterns found in daily DO across a broad range of lakes, and our model training data cover most of the available patterns. Thus, we can expect that our analysis will apply to lakes not in this study if they present patterns that are in the diverse collection herein, which we feel is likely.

Having verified that SAX is appropriate for the globally distributed lakes in our study, it may be appropriate in the future to examine the drivers of differences in DO patterns among lakes, using, for example, environmental and morphological drivers such as season, lake trophic state, climate, location, lake shape, and depth. In addition, the data used in our study is localized to the level of time zone that a lake is within, but not precisely to geographical location. This may have caused a minor inconsistency in the temporal alignment of the patterns, and if a lake's longitude differs from the time zone longitude, a small adjustment to the observation time may be required to apply the model appropriately.

Like all methods, the combined use of expert opinion with SAX transformation has limitations. String similarity discovery models use multiple combinations of letters in a word as the model attribute. For example, in a four-letter word, one can examine the frequency of occurrence of two-letter (e.g., [a][a], [c][b]) or four-letter sequences (e.g., [a][a][c][b]) or the occurrences of two separated letter combinations in the word (e.g., [a][\*][c]). When the size of the words or available alphabet size increases, the number of possible patterns used as model attributes increases exponentially. This results in a need

for greater computational memory and runtime. To limit this from happening, models often deploy n-gram tokenizers that provide a minimum and maximum number of letter combinations for model inputs (Whitelaw et al. 2009). In our study, the SAX letter length and alphabet size for daily DO transformations did not create a major demand on computing resources, as the maximum resource demanded was for SAX(6,6) with  $\sum_{i=1}^6 6^i = 55,986$  predictor variables. For larger datasets, one would require further consideration of the maximum SAX sequences relevant to the frequency and length of the data of interest.

When there was unanimity (Y0 or Y7) among the experts, decisions were generally made without “maybe.” This confirms similar underlying logic that the experts used to determine the data quality. The Y7 label (full consensus; indicating suitable-quality data) was by far the largest populated class, indicating a high occurrence of “textbook quality” data in the observations. The survey also revealed, however, that experts had different expectations about the quality of data. This was evidenced in the large variation between experts in the data that was selected to be suitable (ranging between 34% and 80%). Without a full consensus among the expert panel, it is difficult to make a strong judgment about what is suitable and unsuitable data. It is also unreasonable to disregard any expert's opinion simply because it is in the minority, hence ordinal type expert panel decisions Y0–Y7 were created from the survey instead of a majority decision, to express confidence in the data by the expert panel.

Model performance assessment metrics require careful consideration. For example, for habitat niche distribution model evaluation practices, Lobo et al. (2007) suggest stating the true positive rate (TPR) and true negative rate (TNR) in addition to the AUC values, to further reduce the chance of class imbalance biases. We adapted two different metrics of AUC and MCC together when choosing the model, where the MCC method contains concepts of both TPR and TNR. Both metrics suggested that SAX(4,3) gives the appropriate set of transformation parameters, providing validation in the use of this model. With a training data set of 300 d, higher orders in the SAX transformation, e.g., SAX(6,6), would raise concerns regarding over-fitting of the model. While overfitting may be a concern even with SAX(4,3), the 10-fold cross-validation suggests this complexity of SAX is a reasonable compromise between goodness-of-fit and generalizability. The confusion matrix provided useful insights about the model, and it also identified a few extreme error occurrences. The causes of extreme errors may be due to simple QA/QC type issues (e.g., appearance of partially repeated values and low levels of DO variation; Fig. 7a,c,f), or lack of information to drive the model (e.g., DO increased before sunrise; Fig. 7b,e). While the former QA/QC issue can be easily filtered out, the latter type of error requires additional information to help identify the specific nature of the problem. It should also be noted that the tool identifies repeated values or the occurrence of no data for an entire day as a specific sequence (i.e., aaaa), which is

likely to be classified as Y0. One extreme error instance was not caused by such errors, but was most probably due to simplification made by SAX(4,3) (Fig. 7d). Identification of extreme errors (if they exist) can be helped by comparing class differences between the model outputs of SAX(4,3) and other parameter sets such as SAX(6,6). Nevertheless, the occurrence of such extreme errors related to a lack of information or SAX simplification was minimal (< 1%).

## Conclusions

Hutchinson (1957) wrote that a “skilful limnologist can probably learn more about the nature of a lake from a series of oxygen determinations than from any other kind of chemical data. If these oxygen determinations are accompanied [by additional variables], a very great deal is known about the lake.” This quote illustrates how a highly complex lake system can be evaluated with a series of dissolved oxygen observations, integrating both biogeochemical and physical processes. Most numerical modeling practices involve attempting to capture the majority of these processes using a highly complex set of equations and using a comprehensive dataset (e.g., Robson 2014), but such models themselves demand considerable time, effort and expertise. Theory-guided data science (e.g., Karpatne et al. 2016) tries to capture the essence of the system (i.e., the information contained within the system) by using a less complex model structure, which may be appropriate when the dataset is incomplete. The DO data provided to the classification model in our study effectively represented a minimum level of information (DO shape) but the training dataset was complemented by an expert survey process which involved supplementary data. The use of classes together with additional information commonly requested by the experts (PAR, wind speed, and surface water temperature) is supported by the model performance. The expert survey results tended to confirm that the removal of data was predominantly due to the influence of non-biological processes. Most lake metabolism models assume that biological processes of oxygen production and consumption dominate DO fluxes (e.g., Peeters et al. 2016). However, a variety of non-biological processes can be important in redistributing DO, and these difficult to distinguish phenomena can appear in sensor observations (e.g., Brand et al. 2008). A data mining procedure represents an intermediate level of complexity to capture “suitable” and “unsuitable” observations and the dominant biological and non-biological features of the high-frequency DO sensor data.

The source code can be accessed at: <https://github.com/kohjim/DOClassifier>.

## References

Aggarwal, C. C., A. Hinneburg, and D. A. Keim. 2001. On the surprising behavior of distance metrics in high dimensional space, p. 420–434. *In* V. Den Bussche and Victor Vianu

- [eds.], Database Theory – ICDT 2001. Springer-Berlin. doi: [10.1007/3-540-44503-X\\_27](https://doi.org/10.1007/3-540-44503-X_27)
- Antenucci, J. P., K. M. Tan, H. S. Eikaas, and J. Imberger. 2013. The importance of transport processes and spatial gradients on in situ estimates of lake metabolism. *Hydrobiologia* **700**: 9–21. doi:[10.1007/s10750-012-1212-z](https://doi.org/10.1007/s10750-012-1212-z)
- Batt, R. D., and S. R. Carpenter. 2012. Free-water lake metabolism: Addressing noisy time series with a Kalman filter. *Limnol. Oceanogr.: Methods* **10**: 20–30. doi:[10.4319/lom.2012.10.20](https://doi.org/10.4319/lom.2012.10.20)
- Bradley, A. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **30**: 1145–1159. doi:[10.1016/S0031-3203\(96\)00142-2](https://doi.org/10.1016/S0031-3203(96)00142-2)
- Brand, A., D. F. McGinnis, B. Wehrli, and A. Wüest. 2008. Intermittent oxygen flux from the interior into the bottom boundary of lakes as observed by eddy correlation. *Limnol. Oceanogr.* **53**: 1997–2006. doi:[10.4319/lo.2008.53.5.1997](https://doi.org/10.4319/lo.2008.53.5.1997)
- Campbell, J. L., and others. 2013. Quantity is nothing without quality. *Bioscience* **63**: 574–585. doi:[10.1525/bio.2013.63.7.10](https://doi.org/10.1525/bio.2013.63.7.10)
- Cannata, A., P. Montalto, M. Aliotta, C. Cassisi, A. Pulvirenti, E. Privitera, and D. Patanè. 2011. Clustering and classification of infrasonic events at Mount Etna using pattern recognition techniques. *Geophys. J. Int.* **185**: 253–264. doi: [10.1111/j.1365-246X.2011.04951.x](https://doi.org/10.1111/j.1365-246X.2011.04951.x)
- Cengiz, T. 2011. Periodic structures of Great Lakes levels using wavelet analysis. *J. Hydrol. Hydromech.* **59**: 24–35. doi: [10.2478/v10098-011-0002-z](https://doi.org/10.2478/v10098-011-0002-z)
- Chang, D. W., P. Hobson, M. Burch, and T. F. Lin. 2012. Measurement of cyanobacteria using in-vivo fluoroscopy—effect of cyanobacterial species, pigments, and colonies. *Water Res.* **46**: 5037–5048. doi:[10.1016/j.watres.2012.06.050](https://doi.org/10.1016/j.watres.2012.06.050)
- Cole, J. J., M. L. Pace, S. R. Carpenter, and J. F. Kitchell. 2000. Persistence of net heterotrophy in lakes during nutrient addition and food web manipulations. **45**: 1718–1730. doi: [10.4319/lo.2000.45.8.1718](https://doi.org/10.4319/lo.2000.45.8.1718)
- Cox, K. C., S. G. Eick, and R. J. Brachman. 1997. Brief application description; visual data mining: Recognizing telephone calling fraud. *Data Min. Knowl. Discov.* **1**: 225–231. doi: [10.1023/A:1009740009307](https://doi.org/10.1023/A:1009740009307)
- Cox, T. J. S., T. Maris, K. Soetaert, J. C. Kromkamp, P. Meire, and F. Meysman. 2015. Estimating primary production from oxygen time series: A novel approach in the frequency domain. *Limnol. Oceanogr.: Methods* **13**: 529–552. doi:[10.1002/lom3.10046](https://doi.org/10.1002/lom3.10046)
- Cremona, F., A. Laas, P. Nöges, and T. Nöges. 2014. High-frequency data within a modeling framework: On the benefit of assessing uncertainties of lake metabolism. *Ecol. Model.* **294**: 27–35. doi:[10.1016/j.ecolmodel.2014.09.013](https://doi.org/10.1016/j.ecolmodel.2014.09.013)
- Frank, E., and M. Hall. 2001. A simple approach to ordinal classification, p. 145–156. *In* Luc De Raedt and P. A. Flach [eds.], Proceedings of the 12th European Conference on Machine Learning (EMCL ’01). Springer-Verlag. doi:[10.1007/3-540-44795-4\\_13](https://doi.org/10.1007/3-540-44795-4_13)
- Giling, D. P., and others. 2017. Delving deeper: Metabolic processes in the metalimnion of stratified lakes. *Limnol. Oceanogr.* **62**: 1288–1306. doi:[10.1002/lno.10504](https://doi.org/10.1002/lno.10504)



- Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. 2009. The WEKA data mining software, p. 10. *In* ACM SIGKDD Explorations Newsletter, v. 11.
- Hamilton, D. P., and others. 2015. A global lake ecological observatory network (GLEON) for synthesising high-frequency sensor data for validation of deterministic ecological models. *Inland Waters* **5**: 49–56. doi:[10.5268/IW-5.1.566](https://doi.org/10.5268/IW-5.1.566)
- Hanson, P. C., D. L. Bade, S. R. Carpenter, and T. K. Kratz. 2003. Lake metabolism: Relationships with dissolved organic carbon and phosphorus. *Limnol. Oceanogr.* **48**: 1112–1119. doi:[10.4319/lo.2003.48.3.1112](https://doi.org/10.4319/lo.2003.48.3.1112)
- Hanson, P. C., S. R. Carpenter, N. Kimura, C. Wu, S. P. Cornelius, and T. K. Kratz. 2008. Evaluation of metabolism models for free-water dissolved oxygen methods in lakes. *Limnol. Oceanogr.: Methods* **6**: 454–465. doi:[10.4319/lom.2008.6.454](https://doi.org/10.4319/lom.2008.6.454)
- Horsburgh, J. S., S. L. Reeder, A. S. Jones, and J. Meline. 2015. Open source software for visualization and quality control of continuous hydrologic and water quality sensor data. *Environ. Model. Softw.* **70**: 32–44. doi:[10.1016/j.envsoft.2015.04.002](https://doi.org/10.1016/j.envsoft.2015.04.002)
- Hutchinson, E. 1957. *MLA Hutchinson, G. Evelyn. "A Treatise on Limnology"*, v. 1. John Wiley & Sons, Inc.
- Kara, E. L., and others. 2012. Time-scale dependence in numerical simulations: Assessment of physical, chemical, and biological predictions in a stratified lake at temporal scales of hours to months. *Environ. Model. Softw.* **35**: 104–121. doi:[10.1016/j.envsoft.2012.02.014](https://doi.org/10.1016/j.envsoft.2012.02.014)
- Karpatne, A., and others. 2016. Theory-guided data science: A new paradigm for scientific discovery. arXiv 1–14.
- Keogh, E., K. Chakrabarti, M. Pazzani, and S. Mehrotra. 2001. Dimensionality reduction for fast similarity search in large time series databases. *Knowl. Inf. Syst.* **3**: 263–286. doi:[10.1007/PL00011669](https://doi.org/10.1007/PL00011669)
- Keogh, E., and S. Kasetty. 2002. On the need for time series data mining benchmarks, p. 102. *In* Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD '02.
- Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection, p. 1137–1145. *In* IJCAI'95 Proceedings of the 14th International Joint Conference on Artificial Intelligence.
- Lauster, G. H., P. C. Hanson, and T. K. Kratz. 2006. Gross primary production and respiration differences among littoral and pelagic habitats in northern Wisconsin lakes. *Can. J. Fish. Aquat. Sci.* **63**: 1130–1141. doi:[10.1139/f06-018](https://doi.org/10.1139/f06-018)
- Lin, J., E. Keogh, S. Lonardi, and B. Chiu. 2003. A symbolic representation of time series, with implications for streaming algorithms, p. 2–11. *In* Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. ACM, New York, NY, USA. doi:[10.1145/882082.882086](https://doi.org/10.1145/882082.882086)
- Lin, J., E. Keogh, L. Wei, and S. Lonardi. 2007. Experiencing SAX: A novel symbolic representation of time series. *Data Min. Knowl. Discov.* **15**: 107–144. doi:[10.1007/s10618-007-0064-z](https://doi.org/10.1007/s10618-007-0064-z)
- Lobo, J. M., A. Jiménez-Valverde, and R. Real. 2008. AUC: A misleading measure of the performance of predictive distribution models. *Glob. Ecol. Biogeogr.* **17**: 145–151. doi:[10.1111/j.1466-8238.2007.00358.x](https://doi.org/10.1111/j.1466-8238.2007.00358.x)
- Ma, Y., X. Meng, and S. Wang. 2016. Parallel similarity joins on massive high-dimensional data using MapReduce. *Concurr. Comput. Pract. Exp.* **28**: 166–183. doi:[10.1002/cpe.3663](https://doi.org/10.1002/cpe.3663)
- Matthews, B. W. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *BBA - Protein Struct.* **405**: 442–451. doi:[10.1016/0005-2795\(75\)90109-9](https://doi.org/10.1016/0005-2795(75)90109-9)
- Niennattrakul, V., P. Ruengronghirunya, and C. A. Ratanamahatana. 2010. Exact indexing for massive time series databases under time warping distance. *Data Min. Knowl. Discov.* **21**: 509–541. doi:[10.1007/s10618-010-0165-y](https://doi.org/10.1007/s10618-010-0165-y)
- Peeters, F., D. Atamanchuk, A. Tengberg, J. Encinas-Fernández, and H. Hofmann. 2016. Lake metabolism: Comparison of lake metabolic rates estimated from a diel CO<sub>2</sub>-and the common diel O<sub>2</sub>-technique. *PLoS One* **11**: 1–24. doi:[10.1371/journal.pone.0168393](https://doi.org/10.1371/journal.pone.0168393)
- Rakthanmanon, T., E. J. Keogh, S. Lonardi, and S. Evans. 2011. Time series epenthesis: Clustering time series streams requires ignoring some data, p. 547–556. *In* 2011 IEEE 11th International Conference on Data Mining ICDM.
- Ralanamahatana, C. A., J. Lin, D. Gunopulos, E. Keogh, M. Vlachos, and G. Das. 2005. Mining time series data, p. 1069–1103. *In* O. Maimon and L. Rokach [eds.], *Data mining and knowledge discovery handbook*. Springer, Boston, MA. doi:[10.1007/0-387-25465-X\\_51](https://doi.org/10.1007/0-387-25465-X_51)
- Read, J. S., D. P. Hamilton, I. D. Jones, K. Muraoka, L. A. Winslow, R. Kroiss, C. H. Wu, and E. Gaiser. 2011. Derivation of lake mixing and stratification indices from high-resolution lake buoy data. *Environ. Model. Softw.* **26**: 1325–1336. doi:[10.1016/j.envsoft.2011.05.006](https://doi.org/10.1016/j.envsoft.2011.05.006)
- Richardson, D. C., C. C. Carey, D. A. Bruesewitz, and K. C. Weathers. 2017. Intra- and inter-annual variability in metabolism in an oligotrophic lake. *Aquat. Sci.* **79**: 319–333. doi:[10.1007/s00027-016-0499-7](https://doi.org/10.1007/s00027-016-0499-7)
- Robson, B. J. 2014. State of the art in modelling of phosphorus in aquatic systems: Review, criticisms and commentary. *Environ. Model. Softw.* **61**: 339–359. doi:[10.1016/j.envsoft.2014.01.012](https://doi.org/10.1016/j.envsoft.2014.01.012)
- Rose, K. C., L. A. Winslow, J. S. Read, E. K. Read, C. T. Solomon, R. Adrian, and P. C. Hanson. 2014. Improving the precision of lake ecosystem metabolism estimates by identifying predictors of model uncertainty. *Limnol. Oceanogr.: Methods* **12**: 303–312. doi:[10.4319/lom.2014.12.303](https://doi.org/10.4319/lom.2014.12.303)
- Ruan, G., P. C. Hanson, H. A. Dugan, and B. Plale. 2017. Mining lake time series using symbolic representation. *Ecol. Inform.* **39**: 10–22. doi:[10.1016/j.ecoinf.2017.03.001](https://doi.org/10.1016/j.ecoinf.2017.03.001)
- Schapire, R. E., P. Stone, D. McAllester, M. L. Littman, and J. A. Csirik. 2002. Modeling auction price uncertainty using boosting-based conditional density estimation, p. 546–553.

- In* Proceedings of the Nineteenth International Conference on Machine Learning.
- Solomon, C. T., and others. 2013. Ecosystem respiration: Drivers of daily variability and background respiration in lakes around the globe. *Limnol. Oceanogr.* **58**: 849–866. doi:[10.4319/lo.2013.58.3.0849](https://doi.org/10.4319/lo.2013.58.3.0849)
- Staehr, P. A., D. Bade, M. C. Van de Bogert, G. R. Koch, C. Williamson, P. Hanson, J. J. Cole, and T. Kratz. 2010. Lake metabolism and the diel oxygen technique: State of the science. *Limnol. Oceanogr.: Methods* **8**: 628–644. doi:[10.4319/lom.2010.8.628](https://doi.org/10.4319/lom.2010.8.628)
- Van de Bogert, M. C., S. R. Carpenter, J. J. Cole, and M. L. Pace. 2007. Assessing pelagic and benthic metabolism using free water measurements. *Limnol. Oceanogr.: Methods* **5**: 145–155. doi:[10.4319/lom.2007.5.145](https://doi.org/10.4319/lom.2007.5.145)
- Weathers, K., and others. 2013. The Global Lake Ecological Observatory Network (GLEON): The evolution of grassroots network science. *Limnol. Oceanogr. Bull.* **22**: 71–73. doi:[10.1002/lob.201322371](https://doi.org/10.1002/lob.201322371)
- Whitelaw, C., B. Hutchinson, G. Y. Chung, and G. Ellis. 2009. Using the web for language independent spellchecking and autocorrection, p. 890–899. *In* Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, **2**, EMNLP 2009.
- Winslow, L. A., J. A. Zwart, R. D. Batt, H. A. Duggan, R. I. Woolway, J. R. Corman, P. C. Hanson, and J. S. Read. 2016. LakeMetabolizer: An R package for estimating lake metabolism from free-water oxygen using diverse statistical models. *Int. Waters* **6**: 622–636. doi:[10.1080/IW-6.4.883](https://doi.org/10.1080/IW-6.4.883)
- Witten, I. H., E. Frank, M. A. Hall, and C. Pal. 2016. Data mining: Practical machine learning tools and techniques. Morgan Kaufmann.
- Woolway, R. I., I. D. Jones, D. P. Hamilton, S. C. Maberly, K. Muraoka, J. S. Read, R. L. Smyth, and L. A. Winslow. 2015. Automated calculation of surface energy fluxes with high-frequency lake buoy data. *Environ. Model. Softw.* **70**: 191–198. doi:[10.1016/j.envsoft.2015.04.013](https://doi.org/10.1016/j.envsoft.2015.04.013)
- Zimek, A., E. Schubert, and H.-P. Kriegel. 2012. A survey on unsupervised outlier detection in high-dimensional numerical data. *Stat. Anal. Data Min.* **5**: 363–387. doi:[10.1002/sam.11161](https://doi.org/10.1002/sam.11161)
- ## R Packages
- Chang, W., J. Cheng, J. J. Allaire, Y. Xie, and J. McPherson (2017). shiny: Web Application Framework for R. R package version 1.0.5. Available from <https://CRAN.R-project.org/package=shiny>
- Hornik, K. (2018). RWeKajars: R/Weka Interface Jars. R package version 3.9.2-1. Available from <https://CRAN.R-project.org/package=RWeKajars>
- Hornik, K., C. Buchta, and A. Zeileis. 2009. Open-source machine learning: R meets Weka. *Comput. Stat.* **24**: 225–232. doi:[10.1007/s00180-008-0119-7](https://doi.org/10.1007/s00180-008-0119-7)
- R Core Team (2017). R: A language and environment for statistical computing. R foundation for statistical computing, Vienna Austria. Available from <https://www.R-project.org/>
- Simon Urbanek (2017). rJava: Low-level R to Java Interface. R package version 0.9-9. Available from <https://CRAN.R-project.org/package=rJava>
- ## Acknowledgments
- We thank multiple anonymous experts for providing their insights into data quality, and Chris Solomon and Peter Staehr for letting us use their data products. We also thank the reviewers for their detailed and constructive comments and suggestions. KM thanks Adam Hartland for his support for the manuscript writing. This research was supported by the New Zealand Ministry of Business, Innovation and Employment (UOWX1503; Enhancing the health and resilience of New Zealand lakes). This work benefited from participation in GLEON. The North Temperate Lakes LTER provided data and funding support for PCH.
- ## Conflict of Interest
- None declared.
- Submitted 13 December 2017  
Revised 31 August 2018  
Accepted 04 September 2018
- Associate editor: Clare Reimers