

# **Lake metabolism variability: Identifying physical and biological events using support vector machine**

Paul C. Hanson<sup>1</sup>, Yu Hen Hu<sup>2</sup>, Timothy K. Kratz<sup>3</sup>, Peter Shin<sup>4</sup>

<sup>1</sup>University of Wisconsin – Madison, Center for Limnology  
680 North Park Street  
Madison, Wisconsin 53706

<sup>2</sup>University of Wisconsin – Madison, Electrical and Computer Engineering  
1415 Engineering Drive  
Madison, Wisconsin 53706

<sup>3</sup>University of Wisconsin – Trout Lake Station  
10810 County Highway N  
Boulder Junction, WI 54512

<sup>4</sup>San Diego Supercomputer Center, University of California at San Diego  
MC 0505 9500 Gilman Dr  
La Jolla, California 92037

Running head: Lake metabolism variability

## **Abstract**

Estimates of lake metabolism based on the free-water gas method have been used to assess trophic state in lakes. However, lakes are dynamic systems, and the diel dissolved oxygen (DO) measurements on which the estimates are based can be controlled by biological, as well as physical and chemical processes. Little is understood about variability in surface water metabolism, and there are no established criteria for determining when metabolism estimates are controlled by physico-chemical, as opposed to biological processes. In this study, we investigate lake metabolism variability in the surface waters of north temperate lakes. We use free-water estimates of metabolism from lakes representing a broad range of trophic states to develop heuristics for classifying metabolism estimates as typical (i.e., representing biological processes) or atypical (i.e., representing physical processes). We use the classified data to train a support vector machine (SVM), which is an empirical classifier designed to extract maximum information content from data. The trained SVM adequately captured the “expert knowledge” embodied in the heuristic and outperformed the heuristic in classifying full-season data. In dystrophic and eutrophic lakes, metabolism estimates were more variable and a high percentage (50-80%) of metabolism estimates were classified as atypical. This suggests that physico-chemical processes often disrupt the normal biological control over diel DO.

## Introduction

Lake metabolism is an important indicator of lake biological activity and landscape-lake nutrient fluxes. In lakes, the balance between the carbon fixation through gross primary production (GPP) and mineralization of organic compounds through respiration (R) indicates the relative contributions of autochthonous and allochthonous carbon sources to ecosystem metabolism (del Giorgio and Peters 1994; Cole et al 2000; Hanson et al. 2003). This balance also helps indicate whether lakes are net sources or sinks of atmospheric carbon (Hanson et al. 2004), and the degree to which terrigenous carbon contributes to zooplankton and fish biomass (Cole et al. 2002; Pace et al. 2004). The responsiveness of metabolism to phosphorus loading (Schindler et al. 1978) and the alarming increase in cultural eutrophication of lakes argue for the measurement of metabolism as an indicator of anthropogenic influence on ecosystem condition (NRC 2000). The development of tools to automate metabolism measurement, analysis, and interpretation will help realize its use at broader spatio-temporal scales under a variety of ecosystem conditions.

Diel changes in dissolved oxygen (DO) in lake water can be used to estimate metabolism. With the application of a simple model (Odum 1956), diel DO can provide estimates of GPP, R, and net ecosystem production ( $NEP=GPP-R$ ) (Cole et al. 2000). The model attributes changes in DO to the combined effects of GPP, R, advection (A), and exchange with the atmosphere (F). When the model has been applied to lake data, F has been modeled as a function of wind speed (Cole and Caraco 1998), and A has been assumed to be small (Cole et al. 2000, Hanson et al. 2003). Recent applications of this model have assumed that biology drives the diel DO signal, but others have found that DO can be influenced by non-biological drivers acting at similar temporal scales (Eckert et al. 2003).

Although metabolism estimates made from diel DO have been used to assess the trophic state of lakes, lakes are dynamic systems that periodically undergo rapid change, indicating a change in the biological, chemical, or physical processes driving them. Detecting ecological change as it occurs could allow for intensive study of the ecological driving processes, and integrating change-detection into sensor networks is a goal of the ecological sensing community (Estrin et al. 2003). Can we use metabolism estimates to identify times of rapid ecosystem change? Doing so would require criteria for defining “normal” metabolism conditions that describe ecosystem state, and unfortunately, such criteria do not exist for free-water metabolism estimates. Even more fundamentally, seasonal variability in free-water metabolism estimates has not been documented, preventing us from using descriptive statistics to identify the properties of metabolism distributions.

Discriminating between metabolism estimates that represent ecosystem state from those representing ecosystem change can be posed as a classification problem. In complex systems such as lakes, experts with experience in interpreting DO data perform the metabolism classification post hoc. In ecological sensor networks with high-frequency sampling over broad spatial scales, visually inspecting the data in a timely manner may not be feasible. Furthermore, developing the criteria for determining what is biologically driven versus non-biologically driven change in metabolism for a specific lake may depend on variability statistics derived from full season measurements that may not be available. Advances in machine learning algorithms provide promising techniques for replicating and automating the lake metabolism classification made by experts. Support vector machine (SVM) (Vapnik, 1982; Vapnik, 1995; Vapnik, 1998) is a powerful machine learning algorithm for pattern classification and function approximation.

Theoretically, it offers an optimal solution that minimizes the expected error when a pattern classifier is tested on unseen testing data. Empirically, SVM has been shown to outperform many existing pattern classification algorithms in numerous applications. Furthermore, efficient implementation of the SVM algorithm is available in the public domain.

Our goal was to study lake metabolism variability through the development and implementation of a lake metabolism classifier. We describe seasonal variability in surface water metabolism for three lakes with contrasting trophic states. Based on our knowledge of the controls of DO, we develop a heuristic for labeling metabolism estimates as “normal” , i.e., as describing lake metabolic state, or as “abnormal”, i.e., as describing lake transient conditions. We use the heuristic to label metabolism estimates in a set of lakes covering a broad range in trophic statuses and use the labeled data to train and test an SVM classifier. Not only did the SVM capture the knowledge embedded in the heuristic, but it identified important physical and biological phenomena in the three lakes with seasonal data. The development and application of the SVM informed us of the relationships between the occurrence of ecological change and lake trophic state.

## Methods

We developed and tested a lake metabolism classifier, in the form of a support vector machine (SVM). The SVM was trained on data from lakes covering a broad range in trophic states. The training data were pre-labeled by a simple heuristic model meant to represent a human expert. To test the SVM, we classified metabolism estimates from an oligotrophic, a dystrophic, and a eutrophic lake. We describe the development and testing of the lake metabolism classifier in five general steps. (1) Training and testing data were identified to include a variable to be classified (in this case, metabolism) and a set of predictors, or features, that would help in that classification. (2) Features (similar to predictors) were extracted from the data sets. (3) An initial classification of the training data was performed, using a simple heuristic model. (4) The SVM was trained and validated on the pre-labeled training data. (5) The SVM was tested by classifying a separate, test data set.

### *Training and testing data*

Data for training and testing were from lakes in the Northern Highland Lake District of northern Wisconsin. The training data set had 27 unique lakes, 25 of which were sampled in year 2000 and are described in Hanson et al. (2003). Two lakes, Sparkling Lake and Crystal Bog, were sampled in 2002. Most lakes were sampled more than once, resulting in a total of 57 different deployments. Each deployment lasted a minimum of 2 days, and the mean deployment duration was 3.7 days, resulting in a total of 213 lake-days sampled. DOC concentrations ranged from 1.6 to 24.6 mg L<sup>-1</sup>, and Chl concentrations ranged from 2.5 to 56.9 µg L<sup>-1</sup>. In addition to using the data for training the SVM, we describe the metabolism patterns across lakes by calculating mean and standard deviation of metabolism per lake.

The test data set included three lakes sampled during the summer of 2002. Two of the lakes, Peter Lake and Tuesday Lake, are located in the Upper Peninsula of Michigan on the Notre Dame Environmental Research Center property. The test data differed from the training data in that the durations in the test data were longer; however, their geographic region and trophic statuses fell within the ranges in the training data. Peter Lake was sampled for 99 days during summer stratification. It is a highly productive lake with mean summer total phosphorus concentration of 26 µg L<sup>-1</sup> and Chl concentration of 42 µg L<sup>-1</sup>. Mean summer DOC concentration was low to moderate at 5.8 mg L<sup>-1</sup>. Tuesday Lake is low in productivity and moderate in DOC, with a mean summer TP concentration of 12 µg L<sup>-1</sup>, mean summer Chl of 6.8 µg L<sup>-1</sup>, and mean summer DOC of 8.4 mg L<sup>-1</sup>. It was sampled for 86 days during summer stratification. The third lake in the test set, Trout Bog, is located in Vilas County, Wisconsin, and it was sampled for 58 days during summer stratification. This lake had high mean DOC, moderate TP, and moderate Chl concentration (22 mg L<sup>-1</sup>, 14 µg L<sup>-1</sup>, and 10 µg L<sup>-1</sup>, respectively). We report the mean and standard deviations for both GPP and R for these lakes as a description of the distributions of surface water metabolism estimates made over the summer.

For all the lakes, DO and temperature were sampled from surface waters during summer stratification, using a YSI Model 6250 DO/temperature probe, set to sample at least once every 15 min at 1 m depth. We felt that 15 min samples provided reliable estimates of mean daily metabolism through aggregation of at least 20 samples for both GPP and R (see below). The relationship between sampling frequency and reliability of the daily estimate is outside the scope of this paper. Probe calibration, data correction for drift, and metabolism estimates were conducted as described in Hanson and others (2003).

We chose to use daily metabolism estimates because published methods for calculating metabolism are for daily methods (Hanson et al. 2003). Briefly, night-time change in DO is assumed to be a function of R and oxygen exchange with the atmosphere. Daytime change in DO is assumed to be a function of GPP, R, and atmospheric exchange. We calculated three metabolism values for every calendar day:  $R_1$ , as the mean of individual estimates made from midnight until dawn; GPP as the mean of individual estimates from dawn until dusk;  $R_2$  as the mean of individual estimates from dusk until midnight. Net ecosystem production (NEP) was calculated as the difference between GPP and weighted means of  $R_1$  and  $R_2$ .

### ***Feature extraction***

Feature selection is a critical process in developing the SVM. Features are similar to predictors in regression analysis, and ideally will inform the SVM and be uncorrelated. However, SVM is robust to feature correlation as long as they are a well-discriminating set of features in combination. During the training phase, SVM assigns a weight to each feature, and the weight can be used as a selection criterion for the SVM when a linear kernel is employed (see below). Feature selection should be motivated by the domain needs. In this project, our goal was to classify metabolism data using readily available data sources, but we might have included meteorological data, water currents or temperature profiles had they been available. SVM scales up to handle a large number of features by satisfying the most generalizing conditions.

We chose what we felt were the fewest features required to adequately inform the metabolism classifier models. These three features included DOC and Chl, which are two important drivers of lake metabolism (del Giorgio and Peters 1994; Hanson et al. 2003). DOC also drives mixed layer depth (Snucins and Gunn 2000), and lakes with shallower mixed layers are more susceptible to mixing forced by external drivers. Thus, lakes with high DOC may have DO signals influenced by physical phenomena. Chl concentration is directly related to algal biomass. Lakes with high Chl concentration may have elevated GPP and R values, which may indicate an algal bloom. The third feature was the maximum of the absolute values of GPP,  $R_1$  or  $R_2$  for the day. Absolute values were used because data should not be rejected based solely on its sign (see Results), and when using the magnitude of a metabolism value as a feature vector in a SVM, having all the same signs simplifies the classification.

### ***Metabolism classification by heuristic***

Training the SVM required that the training data be pre-labeled (or classified). Classification of training data sets for complex systems often is performed by an expert in the field – one who can interpret the data based on the subtleties of the context. Naturally, different experts may have different interpretations of the data. Even the same human expert may give inconsistent opinions on the same data set when it is examined at different occasions. To reduce inconsistencies in labeling, we have formulated a heuristic as a substitute for the human expert's classification. The heuristic assumes there are two different kinds of lakes – those with highly variable metabolism and those with low metabolism variability, depending on their Chl and DOC concentrations. Oligotrophic lakes tend to have deep thermoclines and low diel DO amplitudes, thus their metabolism variability might be expected to be low. High DOC lakes have shallow thermoclines and may be subject to weather induced mixing and high metabolism variability. High Chl lakes have high productivity and respiration. In the training data set, we assigned lakes to either high or low variability groups, depending on their DOC and Chl concentrations. The DOC and Chl cutoffs for the group assignments were determined by visually identifying step

changes in plots of metabolism standard deviation versus DOC and Chl. Classifying lakes as typical or atypical within the two groups required that we assign an acceptable threshold for metabolism for the two groups. Metabolism estimates above that threshold were classified as atypical. Rather than assign arbitrary thresholds or define thresholds based on summary statistics from noisy data, we fit the thresholds during the training process described below.

### ***Support Vector Machine***

Support Vector Machine (SVM) is a kernel based machine learning algorithm based on statistical learning theory developed in recent years (Vapnick 1995, Schoelkopf, 1999). SVM is an inductive learning method and is similar to traditional regression models. However, unlike the traditional regression models, SVM seeks the most generalizable model instead of the best fitting model on the training dataset. As such, it promises superior performance when tested with data sets that are not part of the training datasets used to develop the SVM model. In other words, a SVM model trained with historical data is expected to perform well for future data sets.

In its simplest form, SVMs employ optimization algorithms to find the most generalizing model in a two-class pattern classification problem: it locates the boundaries between classes and chooses the one that maximizes the margin that separates the elements of these two classes. This most generalizing dividing boundary is implicitly characterized by a set of “support vectors”, which are the smallest subset of training vectors that uniquely define the boundary. Furthermore, SVMs with non-linear kernels can classify in higher dimension than the original feature space with efficient computation. References and additional information about SVM implementations can be found on-line at : <http://www.support-vector.net/>.

To train and run SVMs, we used a software program, LibSVM (Chang and Lin 2001) available at URL <http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html>. LIBSVM is an integrated software package for support vector classification, (C-SVC, nu-SVC ), regression (epsilon-SVR, nu-SVR) and distribution estimation (one-class SVM ). It supports multi-class classification. The basic algorithm is a simplification of both SMO by Platt (Platt 1998) and SVMLight by Joachims (Joachims 1999). It is also a simplification of the modification 2 of SMO by Keerthi et al (Keerthi, et al 1999). The SVM uses features from the training data (similar to predictors in regression analysis), read from text files, to define an optimal model for predicting binary classifications of that data set (similar to responses in regression analysis) provided by an expert. In this study, the expert was represented by the heuristic model described above.

### ***Training and testing the SVM***

We determined a threshold of metabolism above which lake-days were classified as atypical. These thresholds (GPP or R in  $\text{mg O}_2 \text{ L}^{-1} \text{ d}^{-1}$ ) were termed  $\alpha_{\text{low}}$  for the low variability group and  $\alpha_{\text{high}}$  for the high variability group. We searched for the  $\alpha_{\text{low}}$  and  $\alpha_{\text{high}}$  that minimized the false classifications in the SVM. Although this approach assumed that lakes belonged in one of two groups, it circumvented any assumptions about variability differences between the two groups, essentially allowing the SVM to pick the best thresholds weighting its decision with DOC and Chl. A grid search at an interval of  $0.2 \text{ mg O}_2 \text{ L}^{-1} \text{ d}^{-1}$  was performed across the  $\alpha_{\text{low}}$  and  $\alpha_{\text{high}}$  parameter spaces, and local minima were explored using a minimization searching algorithm (Matlab v.6.5, The Mathworks, Inc.) to find the parameters producing the closest agreement between the heuristic and the SVM. Model validation was performed for every parameter combination used to pre-label the training data by the heuristic model. We validated the SVM using a jack-knife procedure, in which the first feature vector (lake-day plus expert

classifications) was withheld from the training data, and the remaining 212 feature vectors were used to train the SVM, which in turn was used to classify the lake held-out as the test case. There were four possible results for each classification instance, with the assumption that the heuristic classification was correct: (1) SVM correctly classified the test case as typical (true typical, TT); (2) SVM correctly classified the test case as atypical (true atypical, TAT); (3) SVM incorrectly classified the test case as typical (false typical, FT); (4) SVM incorrectly classified the test case as atypical (false atypical, FAT). The procedure was then repeated for each vector in the training matrix. Model performance was evaluated using the following statistics applied to the confusion matrix: Sensitivity= $TT/(TT+FAT)$ ; Specificity= $TAT/(TAT+FT)$ ; Accuracy =  $(CT+CAT)/n$ . The parameter set producing the highest Accuracy, while Sensitivity and Specificity were  $> 0.8$ , was considered best.

As a test, the heuristic and the trained SVM were used to classify the three long-term data sets from Trout Bog, Peter Lake, and Tuesday Lake. By comparing the heuristic and SVM, we determined whether or not the use of SVM was even justified. After all, we could simply accept the heuristic classifications. Classification results from the heuristic and the SVM were graphed along with the original time-series data from the lakes, and their performances evaluated by visually comparing their classifications with events evident in the time-series data.

## Results

### *Patterns in metabolism data*

The distributions of metabolism for the three lakes in the test data (Fig. 1; shows R only) reflect their contrasting trophic statuses and clearly relate to variability evident in the time-series data. In Trout Bog, the lake with the shallowest mixed layer, mean daily R ( $\pm$ STD) ( $1.72 \pm 2.65$  mg O<sub>2</sub> L<sup>-1</sup> d<sup>-1</sup>) was the highest of the three lakes, GPP was lowest ( $0.177 \pm 2.04$  mg O<sub>2</sub> L<sup>-1</sup> d<sup>-1</sup>), resulting in the most negative NEP of all the lakes ( $-1.542$  mg O<sub>2</sub> L<sup>-1</sup> d<sup>-1</sup>). Peter Lake, which had the highest Chl concentration, also had the highest mean daily GPP ( $1.529 \pm 1.05$  mg O<sub>2</sub> L<sup>-1</sup> d<sup>-1</sup>) and the second highest R ( $1.175 \pm 1.14$  mg O<sub>2</sub> L<sup>-1</sup> d<sup>-1</sup>). GPP was in excess of R, resulting in positive NEP ( $0.354$  mg O<sub>2</sub> L<sup>-1</sup> d<sup>-1</sup>). In the oligotrophic lake, Tuesday L., mean daily GPP ( $0.657 \pm 0.437$  mg O<sub>2</sub> L<sup>-1</sup> d<sup>-1</sup>) was second lowest and R ( $0.678 \pm 0.617$  mg O<sub>2</sub> L<sup>-1</sup> d<sup>-1</sup>) was the lowest of the three lakes. NEP for this lake was near zero. A comparison of variability among metabolism estimates from the three lakes shows that the standard deviation about doubles from Tuesday L. to Peter L., and then again from Peter L. to Trout Bog. Typical metabolism could be defined in terms of standard deviation. For example, one standard deviation of GPP estimates in the Tuesday L. encompasses a range of 0.22 to 1.1 mg O<sub>2</sub> L<sup>-1</sup> d<sup>-1</sup>; whereas in highly productive Peter L. one standard deviation includes a range from about 0.5 to 2.5 mg O<sub>2</sub> L<sup>-1</sup> d<sup>-1</sup>. A similar comparison among lakes for R shows a slightly broader range, with Tuesday L. ranging from about 0.06 mg O<sub>2</sub> L<sup>-1</sup> d<sup>-1</sup> to 1.3 mg O<sub>2</sub> L<sup>-1</sup> d<sup>-1</sup> and Trout Bog ranging from  $-0.93$  mg O<sub>2</sub> L<sup>-1</sup> d<sup>-1</sup> to 4.4 mg O<sub>2</sub> L<sup>-1</sup> d<sup>-1</sup>.

One approach to screening these data would be to remove impossible values, namely the negative estimates for GPP and R. For metabolism estimates, process error could be either positive or negative, because it represents patchiness in a sensing field that is heterogeneous in DO. The source (i.e., habitat of origin) of a given patch will determine the DO vector at the sensor as it passes the sensor. For example, if a patch were hypolimnetic in origin, we would expect DO at the sensor to decrease, resulting in a negative metabolism estimate. If it were metalimnetic water, then the DO concentration would be higher if the lake had a metalimnetic Chl peak. Patches originating from littoral zones may differ from pelagic patches in ways not yet identified. Without knowledge of the three dimensional movement of water, nor the DO patchiness within the lake, we felt it prudent to accept both positive and negative metabolism values, and applied the rejection thresholds of  $\alpha$  to the absolute values of mean daily estimates of metabolism.

The typical deployment duration in the training data was 3.7 days, and so the variability in mean daily metabolism for these lakes tended to be much greater than in the test data. The standard deviation in the training data was related to DOC and Chl concentrations. A visual inspection of Fig. 2 shows an increase in metabolism variability at a DOC concentration of about 8 mg L<sup>-1</sup> (Fig. 2, A) and at a Chl concentration of about 15  $\mu$ g L<sup>-1</sup> (Fig. 2, B). At DOC < 8 mg L<sup>-1</sup> and Chl < 15  $\mu$ g L<sup>-1</sup> ( $n=32$ ,  $\overline{DOC} = 4.16$  mg L<sup>-1</sup>,  $\overline{chl} = 5.34$   $\mu$ g L<sup>-1</sup>), mean variability of R was low ( $0.43$  mg O<sub>2</sub> L<sup>-1</sup> d<sup>-1</sup>; STD). When either DOC or Chl were above the aforementioned thresholds ( $n=25$ ,  $\overline{DOC} = 14.13$  mg L<sup>-1</sup>,  $\overline{chl} = 29.57$   $\mu$ g L<sup>-1</sup>), mean variability of R was high ( $2.32$  mg O<sub>2</sub> L<sup>-1</sup> d<sup>-1</sup>; STD). Lakes with DOC < 8 mg L<sup>-1</sup> and Chl < 15  $\mu$ g L<sup>-1</sup> were assigned to the low variability group, while all others were assigned to the high variability group.

### *Metabolism thresholds and SVM training*

The SVM was most accurate when both  $\alpha_{\text{low}}$  and  $\alpha_{\text{high}}$  for the heuristic were  $1.6 \text{ mg O}_2 \text{ L}^{-1} \text{ d}^{-1}$ . With these parameter values, approximately 55% of the cases in the high variability group were classified as AT and approximately 25% of the cases in the low variability group were classified as AT (Fig. 3). Setting  $\alpha_{\text{low}}$  equal to  $\alpha_{\text{high}}$  eliminates categorical distinctions based on DOC and Chl, implying one acceptable range of variability for all metabolism estimates. Any metabolism value outside of that range was classified atypical. Accuracy of the SVM training was 94%, and both sensitivity (1.00) and specificity (0.84) were high.

### ***SVM and heuristic testing***

We classified data from the three test lakes, using the heuristic and the trained SVM, and compared AT classifications with anomalies evident in the graphed data. Our lack of established criteria prevented us from providing a “standard” classification against which the heuristic and SVM could be compared. We present all AT classifications from the heuristic and the SVM for all three time-series. There are four graphs for each lake panel in Figure 4. The top graph shows the DO time-series, and the shaded horizontal bars indicate days highlighted for discussion below. The next two graphs show GPP and R (mean weighted  $R_1$  and  $R_2$ ) for each day (dashed lines are  $\pm 1.6 \text{ mg O}_2 \text{ L}^{-1} \text{ d}^{-1}$ ). The bottom graph indicates the days classified as AT by the heuristic or SVM (labels on Trout Bog panel).

Trout Bog shows moderate GPP and high R as typical metabolism characteristics (Fig. 4, A). Three drastic step changes interrupt the DO time-series near days 178, 192, and 197, the first two of which correspond to mixing events (arrows in Fig. 5). The step changes in DO are reflected in big changes to the metabolism estimates for days 177-179 and for day 192. Both models classified these days as AT. The heuristic classified a total of 42 days as AT, whereas the SVM classified 14 days as AT.

In Peter Lake, algal blooms occurred around days 160 and 180, and then nearly continuously from days 195-245 (Chl in Fig. 5). The DO time-series showed heightened GPP, as well as strong DO supersaturation during those same days. The DO saturation point was  $\leq 8 \text{ mg L}^{-1}$  through this time series. Both models classified as AT those days surrounding 160, 180 and 195 (Fig. 4, B). Both models had a high proportion of days classified as AT.

GPP and R in Tuesday Lake were consistent for most of the time-series, reflecting low GPP and R and DO near saturation (Fig. 4, C). An obvious anomaly in R occurred on day 179, though it was not clear what it represented other than a spike in DO on that day, because the Chl data showed no obvious change (Chl in Fig. 5). Both models classified that day as AT. For SVM, it was the only AT day. The heuristic classified six other days as AT, even though Chl conditions in the lake had not changed and the metabolism estimates were subtly higher on those days.

## Discussion

Discriminating between metabolism estimates that describe the ecosystem state and those that indicate change requires a description of the normal state of metabolism. When the scale of interest encompasses surface waters over the stratified season, an accurate description of metabolism variability for any one lake would require full seasonal data, because episodic physical processes can cause big changes, as demonstrated in our data for Trout Bog (Fig. 5). Unfortunately, those data have not been published for a great variety of lakes, and developing a metabolism classifier based on a narrow range of lakes would greatly limit its applicability to other lakes. An alternative to a lake-specific statistical model is a more generalizable empirical model, such as the SVM described here.

Our challenge was to develop a model that adequately classified the metabolism estimates, using the metabolism estimates themselves and simple rules that assumed metabolism variability related to lake trophic status. Instantiation of rules in a simple heuristic model leads to a simplification of the real world that can restrict information extraction from training data and limit the predictive capability of the model. Alternatively, a classifier based on empiricism, such as SVM, has been proven to maximize the use of information content in training data, but SVM needs labeled (or classified) training data as a guide. Hence, the combined approach of heuristic and SVM in this project. Even though parameters were fit for the heuristic model based on its agreement with the SVM, its classifications of the test data differed from those of the SVM. The heuristic seemed to classify too many days as AT in Trout Bog and Tuesday Lake. This is not surprising, considering that DOC and Chl were effectively removed as criteria from the heuristic by setting  $\alpha_{\text{low}}$  and  $\alpha_{\text{high}}$  equal to each other. In other words, for the heuristic all lakes should have the same metabolism variability. The SVM, on the other hand, did not discard the DOC and Chl features in the data, leading to a classifier that considers DOC and Chl in its classification process. In a practical sense, this would lead to fewer “false positive” classifications in an ecological network designed to detect changes in ecosystem state.

Through the process of developing and testing the SVM, we have enhanced our appreciation of the variable nature of aquatic ecosystems. We summarize the relationships between ecosystem change and lake trophic status by graphing the proportion of atypical days, classified by SVM, versus DOC and Chl (Fig. 6). The training deployments are plotted as open circles and the test lakes as filled circles. The test lakes fall within the training data, with the exception of Peter L. in the DOC graph, for which the AT classifications appear to be driven by Chl instead of DOC. The simple inferences are that the proportion of atypical days increases with increasing DOC and Chl, and that deployments of sondes in low Chl, low DOC lakes will rarely produce atypical metabolism estimates. In lakes with DOC above about  $10 \text{ mg L}^{-1}$ , there may be a high percentage of atypical days – perhaps 20-80%. A similar percentage of atypical days will occur in deployments in lakes with Chl in the range of about  $10\text{-}20 \text{ } \mu\text{g L}^{-1}$ . In highly productive lakes ( $\text{Chl} > 25 \text{ } \mu\text{g L}^{-1}$ ) from 50-100% of the days will be atypical. This is not to say that these metabolism estimates will be driven by physical processes, but rather the metabolism estimates will reflect heightened biological activity relative to that found in the training data set.

Diel DO patterns in lakes change with Chl and DOC concentrations. This adds perspective to the hierarchy control of DO put forth by Eckert et al. (2003), who proposed that control over DO in Lake Kinneret was due to processes acting at different scales, according to the following hierarchy:

$$\text{internal waves} \ll \text{biology} < \text{horizontal advection} \sim \text{vertical advection} \quad (1)$$

We suggest a similar pattern for low DOC and Chl lakes in northern Wisconsin. For high Chl lakes, however, biology may play a more prominent role, at least during the stratified season. For high DOC lakes, the advection terms may be even more prominent than suggested by Eq. 1. A full analysis of the hierarchy of control for DO, or for metabolism estimates, is beyond the scope of this project. A rigorous analysis of the frequency spectra of physical and biological variables from a variety of lakes would improve not only our understanding of the DO signal and metabolism estimates at a given location and depth, but the interactions between physical and biological processes in lakes in general. Experiments in which sensors are distributed around lakes at multiple depths could provide invaluable insights into the contributions of lake habitats to the whole lake metabolism estimate.

Screening metabolism estimates at the daily time scale from a single location in a lake may prove to be one of the easier challenges in using and interpreting the free-water gas method for estimating metabolism. The daily frequency constrains our response time to one day and limits our interpretation of the causes and consequences of these changes. Studying metabolism at shorter time scales will require a better understanding of the controls over DO variability at scales ranging from seconds to hours. The temporal variability in lake metabolism at these scales may relate to spatial heterogeneity in the lake and to lake physics. Making measurements at multiple depths and at multiple locations could provide valuable information about the influence of microstratification, advection, mixing, and biological habitats on whole-lake metabolism. Expanding the sensor network to reach across these spatio-temporal scales, and analyzing and interpreting the torrent of data will require the expertise of limnologists, engineers, and information scientists.

As embedded sensor networks become more pervasive in ecological monitoring, scientists will be confronted with data streams of spatio-temporal extent that exceed the capacity of humans to monitor, analyze, and respond in a timely manner. Signal processing tools, such as SVM, will play an increasingly important role as analytical agents acting on the behalf of scientists. Our goal in developing a metabolism classifier was to create a tool that identifies anomalous metabolism events from a relatively simple and common set of input data and that captures the knowledge implicit in a heuristic model. The SVM metabolism classifier developed here is the first example of a formalized approach to screening metabolism estimates made with the free-water DO method. Because SVM places a premium on being generalizable, our model should transfer well to a variety of lakes we have not yet sampled.

## **Acknowledgements**

We are grateful for the help of M. Van de Bogert as field technician, T. Fountain for help with the SVM, and J. Cole for helpful criticisms of the manuscript. This work was funded by the National Science Foundation, through the North Temperate Lakes LTER program, the Cascade Project, and by the Moore Foundation.

## REFERENCES

- Chih-Chung Change, A. C.-J. L. 2001. LIBSVM: A library for support vector machines. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- Cole, J. J., and N. F. Caraco. 1998. Atmospheric exchange of carbon dioxide in a low-wind oligotrophic lake measured by the addition of SF<sub>6</sub>. *Limnol. Oceanogr.* **43**: 647-656.
- 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. *Limnol. Oceanogr.* **45**: 1718-1730.
- Cole, J. J., S. R. Carpenter, J. F. Kitchell, and M. L. Pace. 2002. Pathways of organic carbon utilization in small lakes: Results from a whole-lake <sup>13</sup>C addition and coupled model. *Limnol. Oceanogr.* **47**: 1664-1676.
- Del Giorgio, P. A., and R. H. Peters. 1994. Patterns in planktonic P:R ratios in lakes: Influence of lake trophy and dissolved organic carbon. *Limnol. Oceanogr.* **39**: 772-787.
- Eckert, W., J. Imberger, and A. Saggio. 2002. Biogeochemical response to physical forcing in the water column of a warm monomictic lake. *Biogeochemistry* **61**: 291-307.
- Estrin, Deborah, William Michener, and Gregory Bonito. Environmental Cyberinfrastructure Needs to Distributed Sensor Networks. A Report from a National Science Foundation Sponsored Workshop, 12-14 August 2003. [http://lternet.edu/sensor\\_report/cyberRforWeb.pdf](http://lternet.edu/sensor_report/cyberRforWeb.pdf).
- Gewin, V. 2002. The state of the planet. *Nature* **417**: 112-113.
- Hamilton, D. P., and S. G. Schladow. 1997. Prediction of water quality in lakes and reservoirs. Part I - Model description. *Ecological Modelling* **96**: 91-110.
- 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.
- Hanson, P. C., A. Pollard, D. L. Bade, K. Predick, S.R. Carpenter, and J. Foley. 2004. A model of carbon evasion and sedimentation in temperate lakes. *Global Change Biology* **10**: 1285-1298.
- T. Joachims, Making large-Scale SVM Learning Practical. *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.
- S.S. Keerthi, S.K. Shevade, C. Bhattacharyya and K.R.K. Murthy. 1999. Improvements to Platt's SMO Algorithm for SVM Classifier Design. Technical Report CD-99-14, Control Division, Dept. of Mechanical and Production Engineering, National University of Singapore, Singapore-119260
- Kratz, T. K., L. A. Deegan, M. E. Harmon, and W. K. Lauenroth. 2003. Ecological variability in space and time: insights gained from the US LTER program. *BioScience* **53**: 57-67.
- Macintyre, S., K. M. Flynn, R. Jellison, and J. R. Romero. 1999. Boundary mixing and nutrient fluxes in Mono Lake, California. *Limnol. Oceanogr.* **44**: 512-529.
- NRC. 2000. *Ecological Indicators for the Nation*. Washington DC: National Academy Press.
- Odum, H. T. 1956. Primary production in flowing waters. *Limnol. Oceanogr.* **1**: 103-117.
- Pace, M. L., J. J. Cole, S. R. Carpenter, J. F. Kitchell, J. R. Hodgson, M. C. Van De Bogert, D. L. Bade, E. S. Kritzberg, and D. Bastviken. 2004. Whole-lake carbon-13 additions reveal terrestrial support of aquatic food webs. *Nature* **427**: 240-243.

- J. Platt, *Fast Training of Support Vector Machines using Sequential Minimal Optimization*, in *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, eds., MIT Press (1998).
- Prairie, Y. T., D. F. Bird, and J. J. Cole. 2002. The summer metabolic balance in the epilimnion of southeastern Quebec lakes. *Limnol. Oceanogr.* **47**: 316-321.
- Saggio, A., and J. Imberger. 1998. Internal wave weather in a stratified lake. *Limnol. Oceanogr.* **43**: 1780-1795.
- Schindler, D. W., E. J. Fee, and T. Ruzsyczynski. 1978. Phosphorus input and its consequences for phytoplankton standing crop and production in the Experimental Lakes Area and in similar lakes. *J. Fish. Res. Board Can.* **35**: 190-196.
- B. Schoelkopf, C.J.C. Burges and A.J. Smola, eds., *Advances in Kernel Methods*, Cambridge MA: MIT Press, 1999.
- Snucins, E., and J. Gunn. 2000. Interannual variation in the thermal structure of clear and colored lakes. *Limnol. Oceanogr.* **45**: 1639-1646.
- Vapnik, V. N. 1982. *Estimation of dependences based on empirical data*. Springer-Verlag.
- . 1995. *The nature of statistical learning theory*. Springer-Verlag.
- . 1998. *Statistical learning theory*. Wiley.
- Wilson M.D.A.-S.R.E.L. , A. S. C. A. 2003. Comparison of support vector machine classification to partial least squares dimension reduction with logistic discrimination of hyperspectral data. *Proceedings of the SPIE - The International Society for Optical Engineering Remote Sensing for Environmental Monitoring, GIS Applications, and Geology II*, 23-26 Sept. 2002 **4886**: 487-497.

## FIGURE CAPTIONS

- Figure 1. Frequency distributions of mean daily R in surface waters of the three test lakes during summer stratification.
- Figure 2. Mean standard deviation of R ( $R_{std}$ ) in the training data set as a function (A) DOC concentration, and (B) Chl concentration. Data were grouped into 2 unit bins to reduce spread in the graph.
- Figure 3. Proportion of cases classified as atypical for the lake-days in the low variability and high variability lake groups. The alpha line intersecting the x axis at 1.6 and the y axis at  $\sim 0.25$  and  $\sim 0.55$  was the threshold identified in M3.
- Figure 4. Classification results for the three test lakes, (A) Trout Bog, (B) Peter L., and (C) Tuesday L. Each panel has DO, R, and GPP (units for R and GPP are  $\text{mg O}_2 \text{ L}^{-1} \text{ d}^{-1}$ ) time-series as the top three graphs. The bottom graph in each panel shows days classified as atypical (AT) by the heuristic model and the SVM (labels in A).
- Figure 5. Time-series data from the three lakes in the test data set. In Trout Bog, disruptions in the  $18^\circ\text{C}$  isotherm (left-pointing arrows in the upper graph) correspond to rapid changes in the DO time-series in Trout Bog. In Peter L., changes in Chl concentration correspond to changes in DO around day 180. After day 200, DO and Chl remain elevated. Tuesday L. has relatively low Chl concentration and shows no rapid changes in DO concentration.
- Figure 6. Proportion of AT classifications in the training data over gradients of (A) DOC and (B) Chl (legend in B). Proportion of AT classifications are graphed for the test lakes, as well, along with their DOC and Chl concentrations ( $\text{mg L}^{-1}$  and  $\mu\text{g L}^{-1}$ , respectively).

Figure 1.

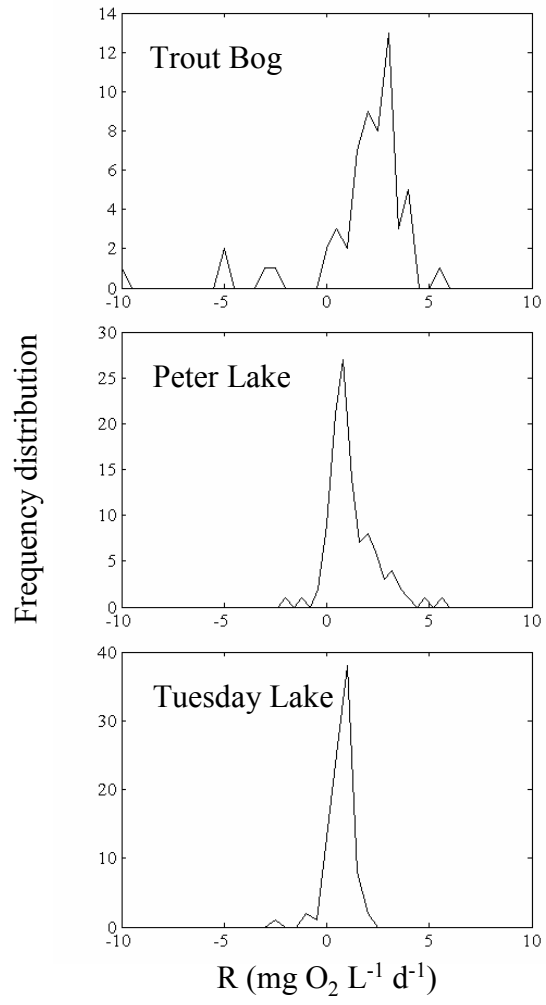


Figure 2.

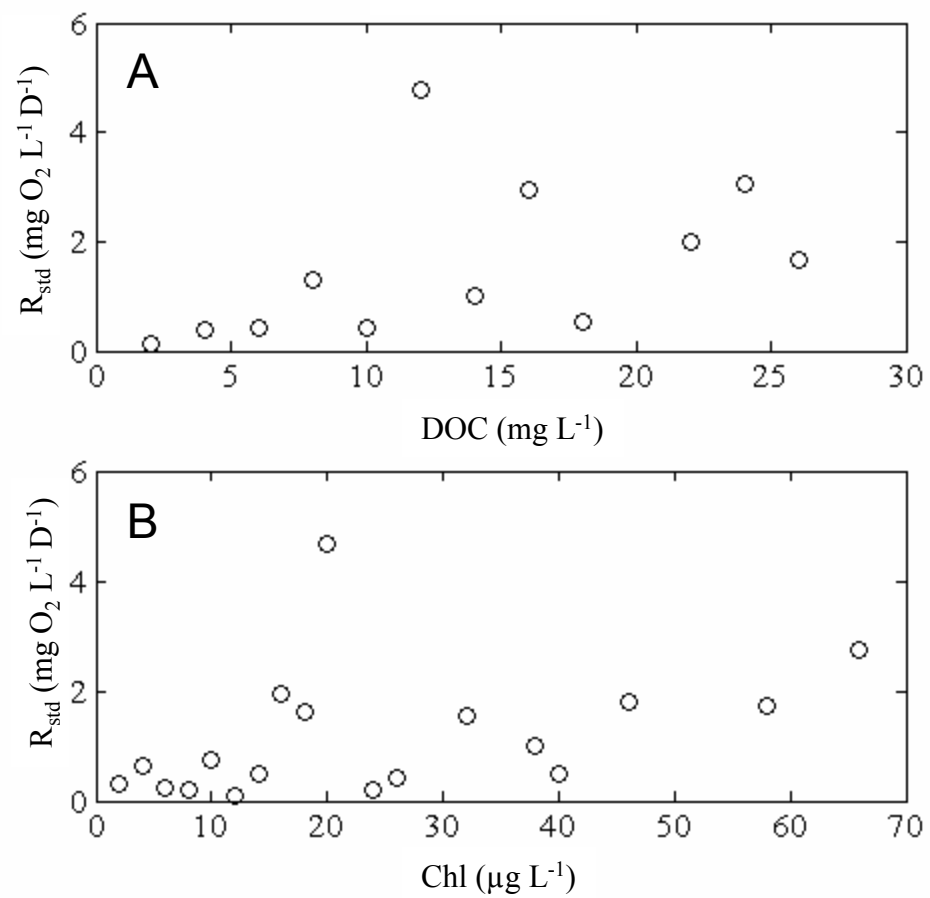


Figure 3.

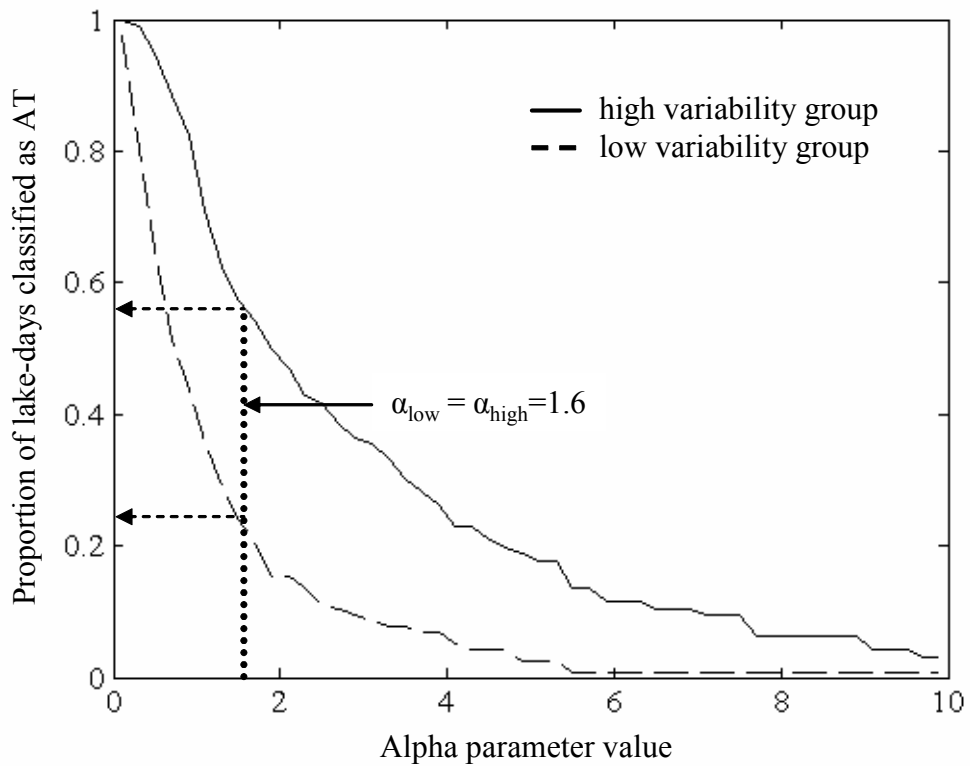


Figure 4.

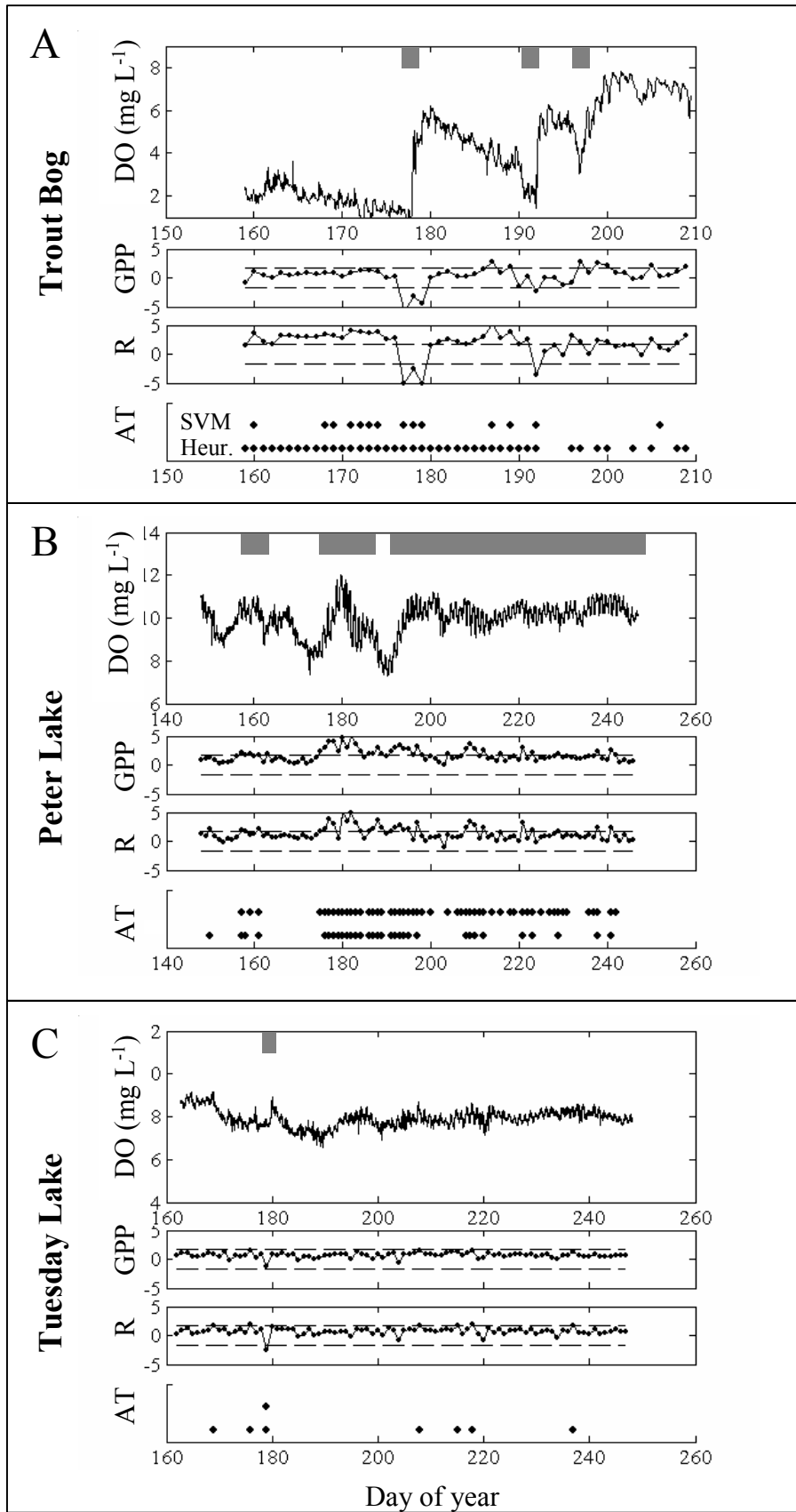


Figure 5.

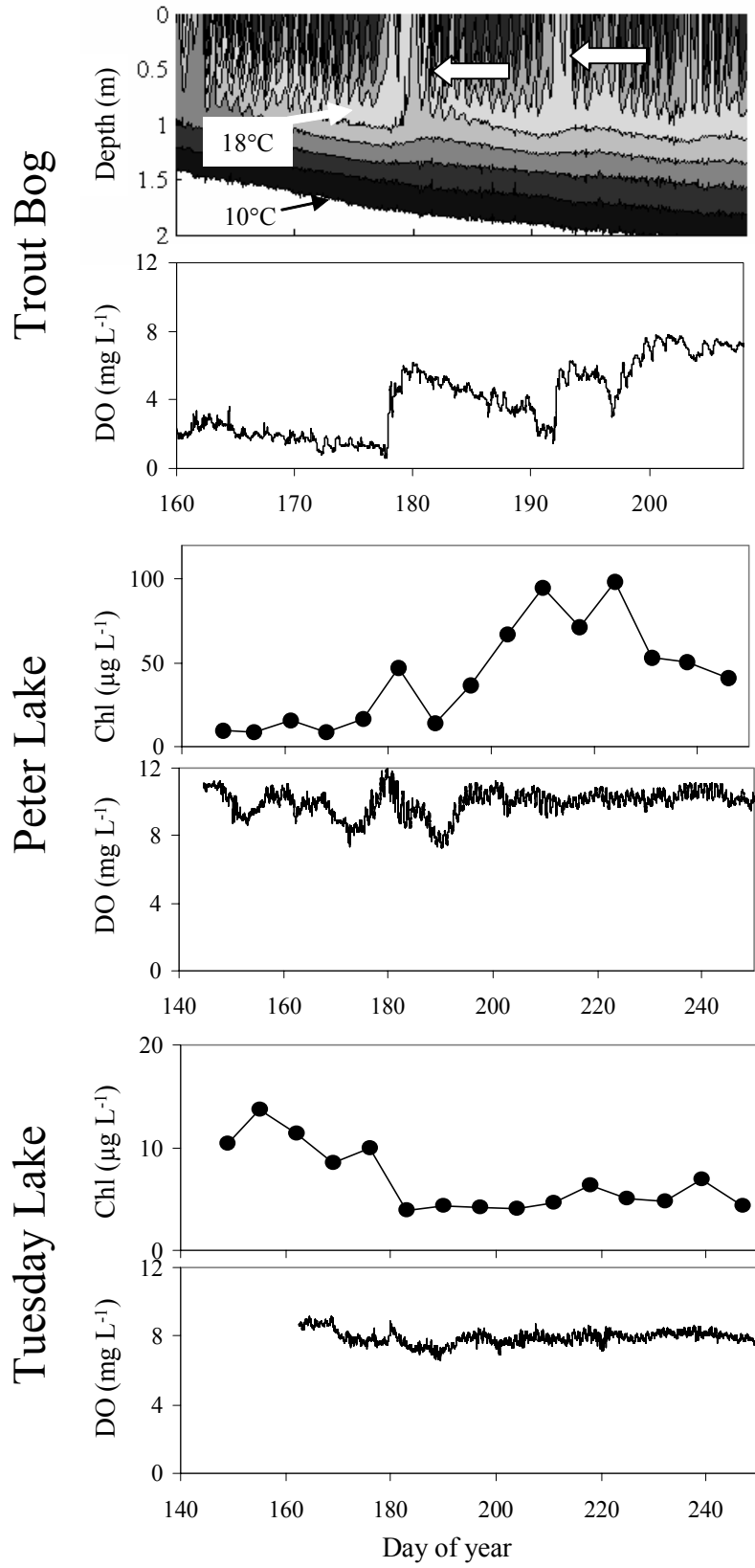


Figure 6.

