Freshwater Biology (2008) **53,** 897–911 doi:10.1111/j.1365-2427.2008.01943.x

Empirical modelling of lake water-temperature relationships: a comparison of approaches

SAPNA SHARMA, STEVEN C. WALKER AND DONALD A. JACKSON
Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, ON, Canada

SUMMARY

1. As a result of the role that temperature plays in many aquatic processes, good predictive models of annual maximum near-surface lake water temperature across large spatial scales are needed, particularly given concerns regarding climate change. Comparisons of suitable modelling approaches are required to determine their relative merit and suitability for providing good predictions of current conditions. We developed models predicting annual maximum near-surface lake water temperatures for lakes across Canada using four statistical approaches: multiple regression, regression tree, artificial neural networks and Bayesian multiple regression.
2. Annual maximum near-surface (from 0 to 2 m) lake water-temperature data were obtained for more than 13 000 lakes and were matched to geographic, climatic, lake morphology, physical habitat and water chemistry data. We modelled 2348 lakes and three subsets thereof encompassing different spatial scales and predictor variables to identify the relative importance of these variables at predicting lake temperature.
3. Although artificial neural networks were marginally better for three of the four data sets, multiple regression was considered to provide the best solution based on the combination of model performance and computational complexity. Climatic variables and date of sampling were the most important variables for predicting water temperature in our models.
4. Lake morphology did not play a substantial role in predicting lake temperature across any of the spatial scales. Maximum near-surface temperatures for Canadian lakes appeared to be dominated by large-scale climatic and geographic patterns, rather than lake-specific variables, such as lake morphology and water chemistry.

Keywords: artificial neural networks, Bayesian multiple regression, multiple regression, predictive models, surface water temperature

Introduction

Many ecological stressors are currently affecting aquatic ecosystems, such as invasions of non-native species and loss of habitat, and there are uncertainties about the extent of future climate change. Water temperature plays an important role in several limnological and biological processes, including ice-cover break up (Anderson, Robertson & Magnuson, 1996), species distribution (Magnuson, Crowder & Medvick, 1979), and the growth and survival of many aquatic organisms, including phytoplankton (Staehr & Sand-Jensen, 2006) and fishes (Christie & Regier, 1988; Magnuson, Meisner & Hill, 1990; Jackson, Peres-Neto & Olden, 2001). Ectothermic organisms have body temperatures approximately equal to the water temperatures and are particularly sensitive to water temperature (Magnuson *et al*., 1990). Water temperature also influences the timing of reproduction, development, growth, mortality, year-class strength and metabolism of most species, including fishes

Correspondence: Sapna Sharma, Department of Ecology and Evolutionary Biology, 25 Harbord Street, University of Toronto, Toronto, ON M5S 3G5, Canada.
E-mail: ssharma@zoo.utoronto.ca

(Shuter & Post, 1990; Tonn, 1990; Brandt *et al*., 2002; Casselman, 2002). Therefore, developing a good statistical approach to model maximum near-surface water temperature is important for comparative studies, particularly in the face of climate change.

Lake temperature is influenced by climate, lake morphology, water chemistry and surrounding topography. Climatic variables such as air temperature (i.e. McCombie, 1959; Arai, 1981; Livingstone & Lotter, 1998; Livingstone & Padisák, 2007) and solar radiation (Kettle *et al*., 2004) have been shown to influence lake thermal characteristics. Aspects of lake morphology, such as surface area (Kettle *et al*., 2004), maximum depth (Kettle *et al*., 2004) and mean depth (Shuter, Schlesinger & Zimmerman, 1983; Snucins & Gunn, 2000; Edmundson & Mazumder, 2002) have also been identified as important predictors of surface water temperature. Large and deep lakes tend to be cooler than lakes that are smaller and shallower, provided that other factors, such as geographic location and climatic conditions, are equal. A study of 60 Alaskan lakes found that lake colour and turbidity were more important in explaining surface water temperature than lake morphology (Edmundson & Mazumder, 2002). Snucins & Gunn (2000) found that dissolved organic carbon was related to surface water temperature in 60 small Ontario lakes (although mean daily air temperature was the most important predictor variable), and less clear lakes (measured as dissolved organic carbon) had higher water temperatures.

A few studies have presented models of maximum near-surface water temperature (e.g. Shuter *et al*., 1980, 1983; Snucins & Gunn, 2000; Edmundson & Mazumder, 2002). However, these studies were generally conducted on a small set of lakes at a small spatial scale. In general, we found that such models did not predict water temperature well in evaluations using large-scale-independent data sets (S. Sharma, unpubl. data). Large-scale climatic and geographic patterns dominate prediction of water temperature at large spatial scales, whereas lake characteristics such as morphology and water chemistry may influence water temperatures at a regional scale (Shuter *et al*., 1983; *sensu* Jackson *et al*., 2001). Therefore, a comparison of statistical approaches and using data sets that vary in their spatial extent is essential to understand how different variables influence predicted maximum near-surface temperature at different spatial scales and to assess the relative merits of the different models.

The objectives of our study were twofold. Firstly, we wanted to identify climatic, morphological, physical and chemical variables that predicted maximum near-surface lake water temperatures most effectively at different spatial scales. The analyses were conducted using four data sets which differed in the number of lakes, the number of predictor variables considered and the spatial distribution of the lakes. The second objective was to evaluate and compare the relative strengths and weaknesses of the statistical approaches in predicting maximum near-surface lake water temperature. Multiple linear regression, a widely used statistical approach, was compared with less commonly used statistical approaches: regression tree, artificial neural networks and Bayesian multiple linear regression.

Methods

Data acquisition

Data describing near-surface water temperature (temperatures collected between 0 and 2 m) and corresponding lake morphology, water chemistry and climatic variables were gathered from a variety of academic and government institutions across Canada and from numerous publications and theses. We obtained data for 47 609 Canadian lakes of which 13 072 lakes had information on near-surface water temperature. We included only one water-temperature value for each lake, selecting the value that was the maximum annual near-surface water-temperature recorded for that lake. Variables for which we obtained data included: latitude, longitude, surface area, volume, maximum depth, mean depth, shoreline perimeter, altitude, water temperature (surface or near-surface between 0 and 2 m), water-temperature measurement depth, conductivity, Secchi depth, total phosphorous concentration, total dissolved solids concentration, pH, dissolved oxygen concentration and sampling date (year and the day of year when near-surface water temperature was measured, to account for the intra- and inter-annual variability in maximum near-surface water temperature). Climatic variables were obtained from the IPCC Data Distribution Centre as 1961–1990 averages. The data were provided as interpolations from meteorological

stations using thin-plate splines and summarized on a $0.5° \times 0.5°$ grid. Climatic variables included: mean annual air temperature, mean July air temperature, monthly and mean annual precipitation, monthly and mean annual solar radiation, and monthly and mean annual cloud cover percentages. We calculated the maximum number of daylight hours for each month and an annual mean using tables provided by the U.S. Navy (http://aa.usno.navy.mul/data/docs/Dur_OneYear.html). Further details regarding data acquisition and summary statistics of the data can be found in Sharma *et al*. (2007).

To remove some inherent biases in spatial coverage and missing values in the data, the data set were pruned to 2348 lakes. We only included data collected after 1960 between June and mid-September. The Ontario and Nova Scotia data sets were considerably larger than the other provinces; therefore, we only included lakes that were sampled in July and August as the temperatures attained from these lakes were likely to be closer to the maximum annual near-surface water temperature. Lakes from Ontario and Nova Scotia were subsampled randomly while stratifying for geographical distribution and lake morphology.

Four different data sets were constructed that differed in the number and types of predictor variables, number of lakes and the spatial extent to determine the relative importance of different predictor variables at varying geographic scales (maximum distance apart ranging from approximately 1500 to 5050 km). Variables selected were included based on the their ecological relevance as identified in the literature. Several factors were taken into consideration when determining the inclusion of predictor variables in the four data sets: spatial extent, sample size and multicollinearity. The four-predictor model was constructed such that the spatial coverage spanned most of Canada and thus represented the broad-scale maximum lake near-surface water-temperature model. The four-predictor model also comprised the largest sample size and the variables with the lowest variance inflation factor (indicating the least amount of collinearity between predictor variables). The spatial coverage and the sample size of the nine-predictor model was smaller, but also included information on lake morphology and water chemistry. This data set covered lakes in Alberta, Ontario, Nova Scotia and Newfoundland. The 10-predictor model also included the depth at which water temperature was recorded, but this further reduced spatial coverage and sample size, containing lakes in Ontario and Nova Scotia. The final 17-predictor model data set incorporated information on climate, lake morphology, water chemistry and sampling, but covered only a subset of lakes in Ontario. Therefore, the division of the four data sets allowed us to assess the effect of spatial scale on the role of climate, lake morphology and water chemistry in predicting maximum near-surface temperature. The data sets comprising the four-, nine-, 10- and 17-predictor variables were divided randomly into training and validation data sets based on a 70 : 30 split (Table 1; Fig. 1) to provide independent evaluations of the predictive abilities of each resulting model.

Data analyses

Multicollinearity between predictor variables were evaluated using bivariate plots, correlation coefficients and variance inflation factors to determine which variables to retain when collinearity was an issue. Variables were transformed as necessary to satisfy the assumption of normality and homoscedacity. Log-transformed variables were: surface area, maximum depth, mean depth, perimeter, altitude and total dissolved solids. Four statistical approaches (linear multiple regression, regression tree, artificial neural networks and Bayesian linear multiple regression) were used.

Statistical approaches used to model continuous response variables

Stepwise multiple-regression models based on forward selection were conducted in SAS. Multiple regression assumes a linear relationship between the response variable and the predictors (Lek *et al*., 1996). The multiple-regression models included year as a variable to account for annual variations, i.e. year represents a non-ordered categorical variable in the analysis. The best stepwise multiple-regression model for each set of maximum number of potential predictor variables based on the Akaike Information Criterion (AIC) was evaluated using the independent, validation data sets. The use of independent, validation data sets provides a more realistic estimate of model prediction error relative to the more traditional

Table 1 Sample sizes for the training and validation data sets and parameters used in the four data sets with four, nine, 10 and 17 predictor variables

Model (km)	$n_{Training}$	$n_{Validation}$	Parameters
Four predictors (*c.* 5050)	1476	872	Mean annual air temperature (mat), mean July air temperature (mjt), year and day of year (doy)
Nine predictors (*c.* 4450)	636	519	mat, mjt, Secchi depth, surface area (SA), doy, June daylength (Juneday), maximum depth (Zmax), mean depth (Zmean), year
10 predictors (*c.* 3725)	500	438	mat, mjt, Secchi depth, SA, Zmax, Zmean, year, doy, Juneday, measurement depth (Zmeas)
17 predictors (*c.* 1500)	141	127	mat, mjt, Secchi depth, SA, Zmax, Zmean, year, Zmeas, doy, Juneday, pH, altitude, total dissolved solids (TDS), mean July precipitation (pptJuly), mean annual precipitation (meanppt), solar radiation in July (Julyrad), July % cloud cover (Julycloud)

The approximate maximum pairwise Euclidean distance between sites is summarized in kilometres for each data set.

resubstitution approach (i.e. bootstrapping, jack-knifing and leave-one-out approaches) because an independent, validation data set tests the model using data not used in its construction (Olden & Jackson, 2000; Sharma & Jackson, in press).

Regression trees can be calculated using both continuous and categorical predictor variables and aim to divide data iteratively into two homogenous groups that have mutually exclusive memberships while maximizing the homogeneity within the two groups (Rejwan *et al.*, 1999; De'ath & Fabricius, 2000; De'ath, 2002). We performed and validated regression trees in SAS. The significance level was set at $P < 0.05$ and the algorithm attempted to minimize the root-mean-square-error at each split. Each tree was pruned to an appropriate size based on cross-validation using the cost-complexity and reduced-error pruning tools available in SAS with the training data set and tested on the independent, validation data set. Cross-validation is the preferred method as the final tree attained using the cross-validation approach tends to have the lowest predicted mean-square-error and should give the most accurate prediction (De'ath, 2002). We used the assessment plot to identify the number of leaves required to significantly reduce the root-mean-square-error. This produced regression trees that were still large enough to isolate rare events. Further details on regression trees can be obtained from De'ath & Fabricius (2000).

Artificial neural networks are designed to mimic the learning process of the mammalian brain and provide a machine-learning approach to minimize some measure of error. The influence of predictor variables (input neurons) is modified and mediated through their connections to a series of hidden neurons, which in turn are connected to the response variable (output neuron). The various pathways by which predictor variables can be linked to the response variable provide the potential for various interactions between variables and nonlinear relationships between predictors and the response variable. Artificial neural networks were based on a single hidden-layer, feedforward, back-propagation procedure in Statistica. The number of hidden neurons was evaluated ranging from one up to the number of predictor variables. Final model selection was based on choosing the model that had the lowest root-mean-square-error for the training data set. Further details on artificial neural networks are available in Olden & Jackson (2002a,b).

Many approaches to Bayesian variable selection are available (e.g. Kuo & Mallick, 1998; Casella & Moreno, 2006; Lunn, Whittaker & Best, 2006). We used the simple and flexible multiple-regression approach of Kuo & Mallick (1998). This approach has been found to work well for several problems in statistical genetics (e.g. Uimari & Hoeschele, 1997). The method of Kuo & Mallick (1998), like other Bayesian variable selection methods, is based on the posterior probability (or probability given the data) that a given subset of predictor variables is the best subset of the set of predictors considered in the analysis. To calculate posterior probabilities, prior probabilities (or probabilities before data analysis) must be specified. In this paper, the prior probability that any predictor variable is included in the model is equal to the prior probability that it is excluded. This objective prior distribution avoids having our prior beliefs influence the predictions. We implemented the method of Kuo & Mallick (1998) in WINBUGS. WINBUGS is a freely

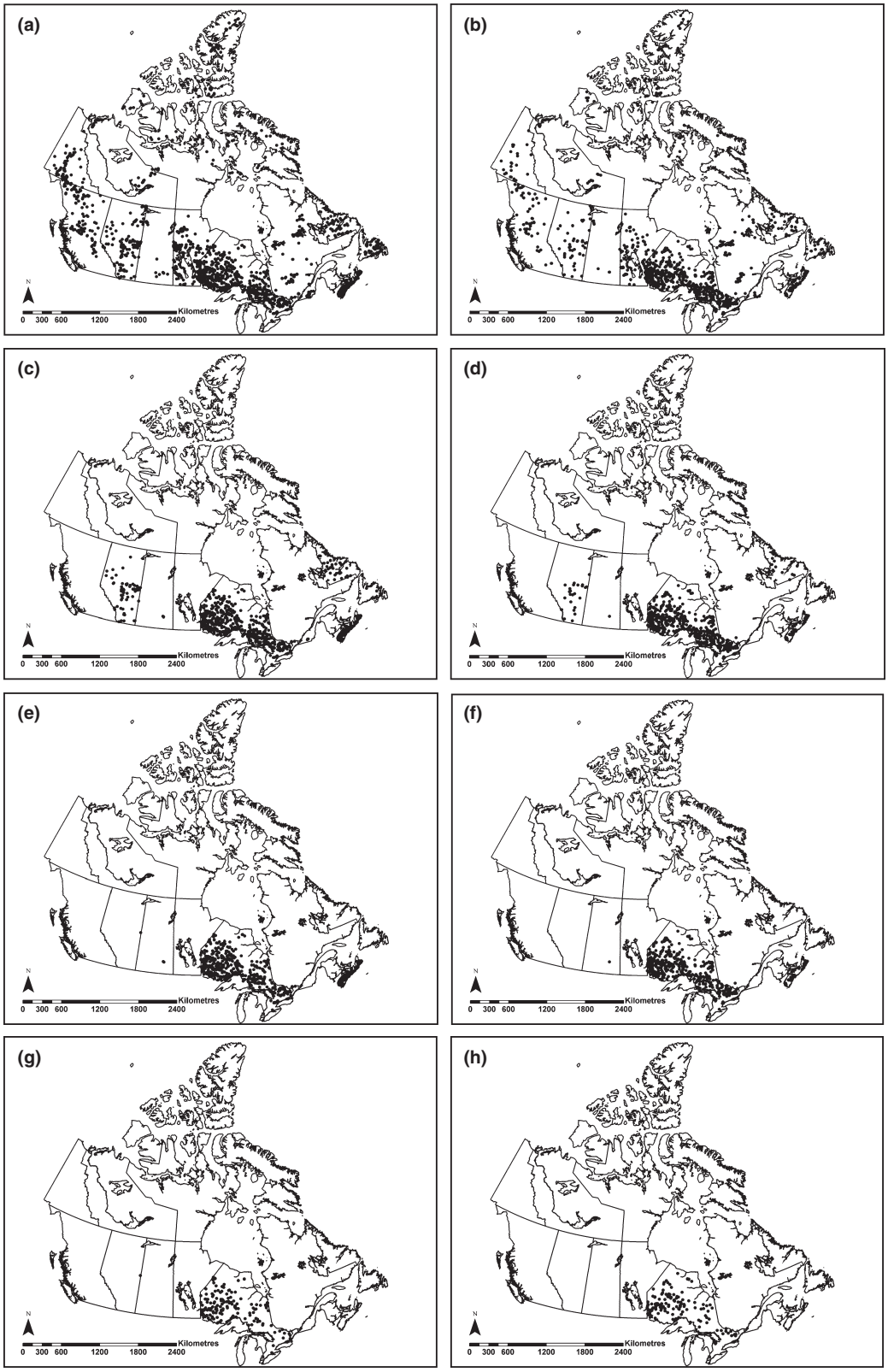

Fig. 1 Location of Canadian lakes comprising training (a, c, e and g) and validation data sets (b, d, f and h) for the four, nine, 10, and 17 predictor variable data sets.

available software package for conducting Bayesian analyses using Markov chain Monte Carlo numerical methods. See Appendix S1 for details on our use of the Kuo & Mallick (1998) approach. We refer readers to the paper by Ntzoufras (2002) for further discussion of Bayesian variable selection methods in WINBUGS.

All models were evaluated on an independent, validation data set. Adjusted R^2, root-mean-square error (RMSE) and the median deviation between observed and predicted values were calculated to evaluate the models. Medians provide a more robust estimate of the fit given that other measures are more influenced by extreme values (Chen & Jackson, 1995). Scatterplots were used to compare observed and predicted maximum near-surface water temperatures for each of the statistical approaches and the four data sets consists of four, nine, 10 and 17 predictor variables. This comparison was conducted to match the relative predictive abilities of each model on the set of independent lakes not used to construct the models.

Results

Across all data sets, multiple-regression analyses revealed that mean July air temperature and year (the year during which each water temperature was recorded) were the predictor variables that generally explained the most variation in maximum near-surface water temperature. Day of year (the date when each water temperature was recorded) and mean annual air temperature were significant predictor variables, but did not account for much variation in water temperature. Lake morphology variables, specifically mean depth and maximum depth, appeared as significant predictors only in the nine-predictor data set and explained <1% of the variation in water temperature. Water chemistry variables did not play a significant role in predicting temperature among the four data sets at any spatial scale. The largest amount of variation explained was in the four-predictor model that explained 77% of the variation in water temperature, but only 34% for the 17-variable model. The RMSE of the models with the validation data set were all <2.7, indicating that the 'average' error associated with any given predicted data set would be <2.7 °C. The median deviation ranged from −0.21 to 0.10 °C when the models were validated on the independent data set (Table 2). The coefficients generated by the multiple-regression models are presented in Table 3.

Mean July air temperature, mean annual air temperature, year and day of year tended to be the most important predictor variables based on the regression tree modelling approach for the four data sets. Lake morphology or water chemistry variables did not appear in any of the regression trees as significant. The four-predictor model accounted for 83% of the variation in near-surface water temperature, compared with only 29% for the 17-variable model. The median deviation ranged from −0.19 to 0.36 °C when the models were validated on the independent data sets. However, the RMSE of the models with the validation data sets ranged from approximately 4.99 to 9.01, indicating considerable error in the predictions produced. These results showed that the models generated by the regression trees were not good predictors of near-surface water temperature on completely independent data sets (Table 2).

Mean July air temperature, year, day of year and June day length (maximum number of daylight hours in June) tended to be the most important variables predicting near-surface water temperatures based on artificial neural networks. Maximum depth and mean depth were morphological variables that were categorized as significant predictor variables in at least one of the models, although they were not the most important. The four-predictor model accounted for 80% of the variation in maximum near-surface water temperature, while the 17-predictor model accounted for only 21%. The median deviation ranged from −0.16 to 0.14 °C when the models were validated on the independent data sets. The RMSE was low for three of the models but for the 17-predictor model on the validation data set the RMSE was 3.86, indicating relatively poor predictive ability compared with some of the other methods. This indicates that, for this particular case, the model was well calibrated for the training data set but this predictive ability did not extend to the independent data set, probably as a consequence of the low number of observations relative to variables (i.e. an overfitted model). Therefore, the models generated by the artificial neural networks were good predictors of maximum near-surface water temperature in the independent data sets, with the exception of the 17-predictor model (Table 2).

Table 2 Evaluation of the four statistical approaches on the (a) four, (b) the nine, (c) the 10 and (d) the 17 predictor variables model and evaluated on the independent, validation data sets

(a)

Statistical approach	Predictor variables retained	Adjusted R^2	RMSE	Median deviation
Multiple regression	mjt, year, doy, mat	0.77	2.66	−0.21
Regression tree	mjt, mat, year, doy	0.83	6.67	−0.06
Artificial neural networks	mjt, year, doy, mat	0.8	2.37	0.05
Bayesian model selection	mat, mjt, doy	0.7	2.88	0.29

(b)

Model	Predictor variables retained	Adjusted R^2	RMSE	Median deviation
Multiple regression	mjt, year, mat, doy, Zmean, Zmax	0.57	2.42	−0.10
Regression tree	mjt, mat, year, doy	0.58	5.98	0.36
Artificial neural networks	doy, year, mjt, mat, Juneday, Zmax	0.59	2.08	0.14
Bayesian model selection	mat, mjt, Zmax, doy	0.35	2.60	0.13

(c)

Model	Predictor variables retained	Adjusted R^2	RMSE	Median deviation
Multiple regression	year, mjt, Juneday, doy, Zmeas	0.39	2.29	−0.09
Regression tree	mat, year, doy, Juneday, Zmeas	0.41	4.99	−0.09
Artificial neural networks	year, doy, Juneday, mjt, Zmeas, mat, Zmax, Zmean	0.46	2.23	−0.16
Bayesian model selection	mjt, Juneday, doy	0.35	2.45	−0.02

(d)

Model	Predictor variables retained	Adjusted R^2	RMSE	Median deviation
Multiple regression	mat, year, pH, doy, altitude, Zmeas	0.34	2.35	0.10
Regression tree	doy, mat	0.29	9.01	−0.19
Artificial Neural Networks	year, Julycloud, Julyrad, pH, TDS, mjt, mat, altitude, doy, mean ppt, Zmean, secchi, July ppt, SA	0.21	3.86	−0.01
Bayesian Model Selection	mat	0.30	2.71	−0.04

RMSE, root-mean-square error.

Mean annual air temperature, mean July air temperature and day of year were the most important predictor variables explaining maximum near-surface water temperatures across the four data sets based on the Bayesian multiple-regression approach. Maximum lake depth was the only morphological variable that appeared as significant in the nine-predictor variable model. Bayesian multiple regression selected the smallest number of predictors for all data sets. This indicates that the Bayesian approach was conservative with respect to including predictors in the models. The four-predictor model accounted for 70% of the variation in near-surface water temperatures, whereas the other models accounted for only 30 or 35% of the variation. The RMSE ranged from 2.45 to 2.88 and the median deviation ranged from −0.04 to 0.29 °C when the models were evaluated on the independent, validation data set (Table 2).

Scatter plot summaries of the predictive modelling approaches were used to assess how they matched up with each another and the observed data. Scatterplots for which the data fall entirely on the diagonal of the plot indicate that the temperatures are identical (Fig. 2). For the data set consisting of four predictor variables, all four statistical approaches matched well with the observed data. Furthermore, near-surface water temperatures predicted by multiple regression and artificial neural networks agreed closely. Predictions generated by the regression tree and Bayesian multiple regression appeared to be most inconsistent with the observed data (Fig. 2a). Similar patterns relative to the four-predictor variable data set (albeit

Table 3 Coefficients for multiple-regression models predicting maximum near-surface water temperatures

Variable	4 predictors	9 predictors	10 predictors	17 predictors
Intercept	7.95	13.27	41.41	8.44
Mean July air temperature	0.81	0.6	0.39	–
Mean annual air temperature	0.23	0.23	–	0.44
Day of year	−0.02	−0.03	−0.02	−0.01
Secchi depth		–	–	–
log Surface area		–	–	–
June daylength		–	−1.59	–
log Maximum depth		2.06	–	–
log Mean depth		−2.08	–	–
Measurement depth			0.06	–
pH				0.81
log Altitude				2.58
log Total dissolved solids				–
Mean July precipitation				–
Mean annual precipitation				–
July solar radiation				–
July % cloud cover				–

The coefficients are not presented for the year term as it is a categorical variable.

less robust relationships) were found in the data sets containing nine- and 10-predictor variables, but with increased prediction errors using the regression tree approach. Predictions generated by multiple regression and Bayesian multiple regression were overestimating low-observed values and under-estimating high-observed values. Predictions from the 10 predictor variable data set using all four modelling approaches also revealed errors. Predictions generated by regression trees and artificial neural networks tended to over-estimate low-water temperatures. Conversely predictions generated by multiple regression and Bayesian multiple regression tended to under-estimate high near-surface water temperatures (Fig. 2b – upper triangle). For the 17 predictor variable data set, the four statistical approaches over-estimated low near-surface water temperatures. The greatest amount of prediction errors were in the 17 predictor variable data set across statistical approaches (Fig. 2b – lower triangle).

Discussion

Statistical methodology

Statistical approaches vary in their response to different data sets depending upon the structural properties of the data. A statistical evaluation of a variety of predictive modelling methods can determine which statistical approach is the most appropriate (Guisan & Zimmermann, 2000). Stepwise multiple regression is traditionally the most popular statistical approach to use with continuous data. Non-traditional approaches such as regression trees, artificial neural networks and Bayesian multiple regression are promising and were evaluated here to determine their ability to predict continuous data. We used adjusted R^2, RMSE and the median deviation from observed and predicted values to evaluate models and to provide an estimate of their relative strengths and weaknesses.

Overall, multiple regression performed very well in predicting maximum near-surface water temperatures in the independent, validation data sets. Multiple regression is one of the most widely used statistical approaches in ecology and is computationally simple relative to the other methods considered. The multiple-regression model assumes that the average of the dependent variable is related linearly to the other variables and it performs very well when that is the case. Whittington *et al.* (2006) outlined several problems with stepwise multiple regression, particularly with data sets containing highly correlated variables. These included biases in the estimation of predictor variables, overfitting of the data, overinflation of R^2, the generation of only one 'best' model, problems with the algorithms used in the analyses, and the problem of testing multiple hypotheses. They suggested the use of the information theoretic based on the AIC to reduce biases. We used AIC and evaluated the model on independent, validation data sets to eliminate potential biases in our study. We did not use a global model (i.e.

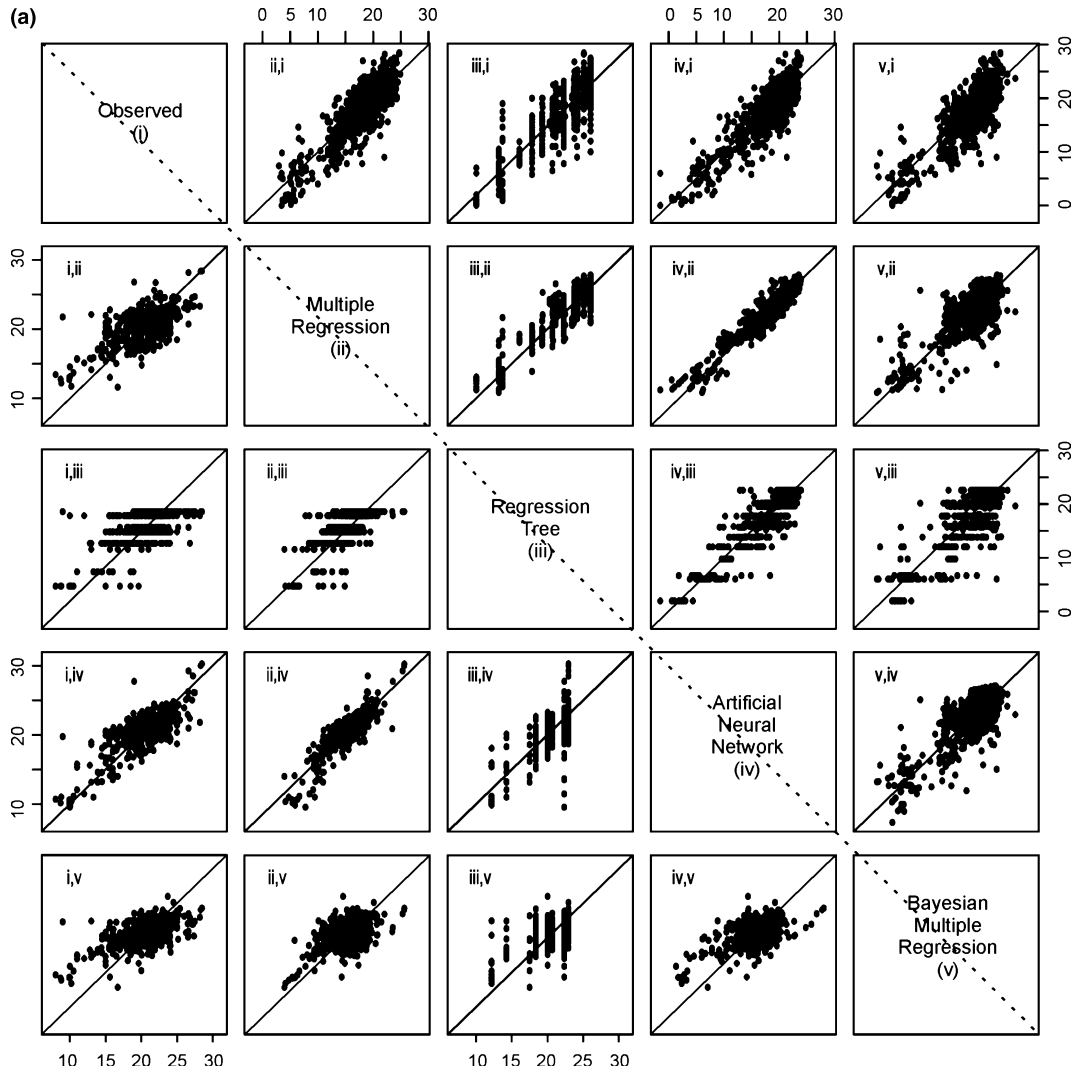

Fig. 2 Scatter plots representing comparisons between observed maximum near-surface water temperatures (i) in the validation data sets and predicted water temperatures based on multiple regression (ii), regression tree (iii), artificial neural networks (iv) and Bayesian multiple regression (v). The 1 : 1 line is included in each scatter plot. The upper right and lower left series of plots are separated by the dotted line in each of panels (a) and (b). The first Roman numeral (i.e. i, ii, iii, iv or v) in each label indicates the values being plotted on the horizontal axis and the second letter indicates the values plotted on the vertical axis throughout the panel. For example, the plot 'i, ii' would present the observed temperature values along the horizontal axis versus the vertical axis being the predicted temperature values from the multiple-regression model. In (a), the upper right part of the panel presents predictions based on the four predictor variable data set (e.g. top, left plot in that panel shows predicted values from the multiple regression model versus observed temperatures, i.e. plot 'ii, i') and the lower left part presents temperature predictions based on the nine predictor variable data set. In (b), the upper right part of the panel presents predictions on the 10 predictor variable data set and the lower part presents predictions on the seventeen predictor variable data set.

deriving the parameters with all of the predictors present) because it can generate excess noise and does not clearly identify the importance of predictor variables (Whittington *et al.*, 2006).

Artificial neural networks generally had high predictive abilities, with the exception of the 17-predictor variable data set for which the model was over-fitted. Artificial neural networks do not require data that meet standard statistical distributions and are appropriate when the underlying distribution of data are complex or unknown and if variables exhibit multicollinearity. Additionally, they are robust to nonlinearity and can model simultaneously multiple predictor variables and their interactions without *a*

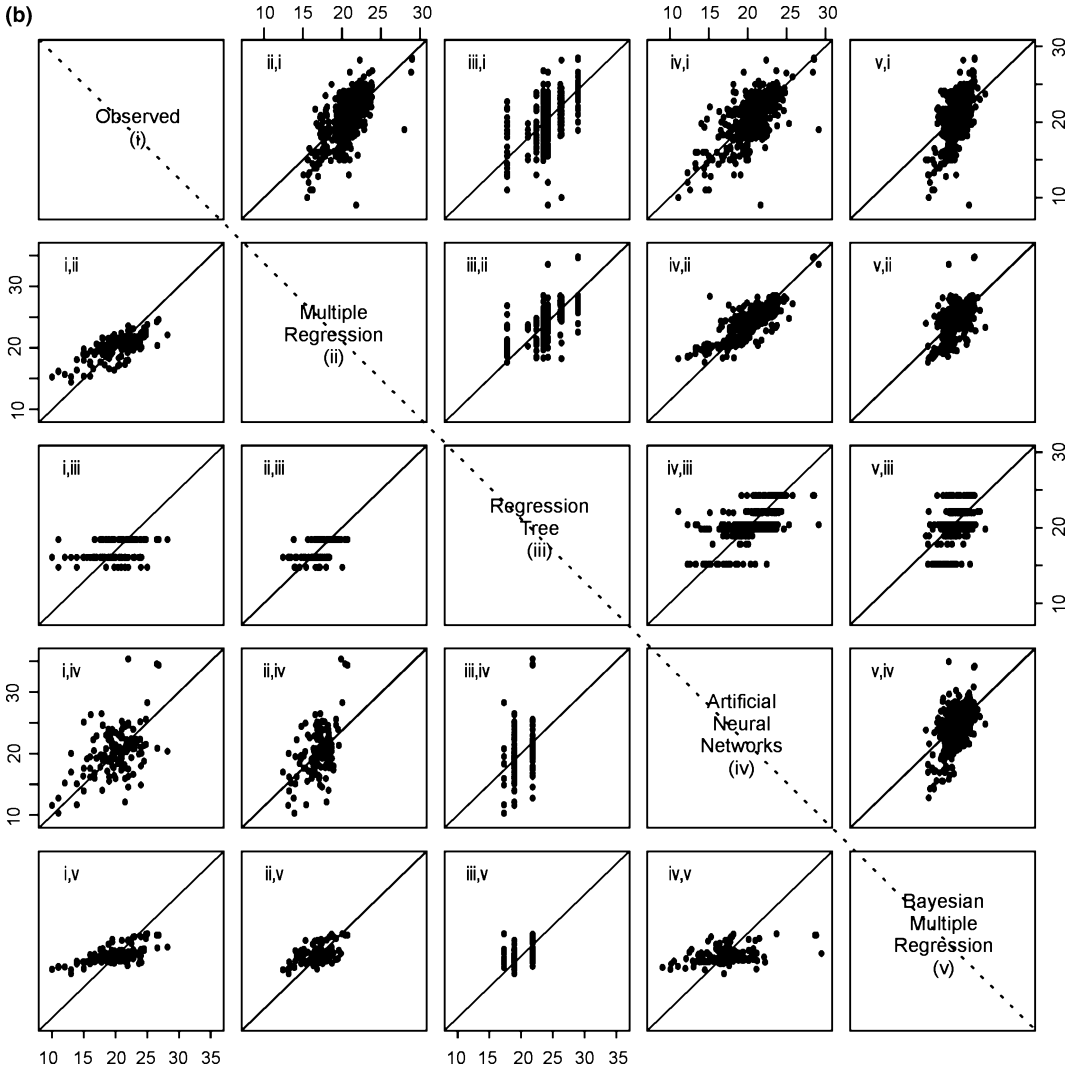

Fig. 2 (*Continued*).

priori knowledge and specification (Brosse *et al*., 1999; Pearce & Ferrier, 2000; Olden & Jackson, 2001; Gevrey, Dimopoulos & Lek, 2006; Olden, Joy & Death, 2006). As we found with the 17-predictor variable data set, artificial neural networks can encounter problems of overtraining (Ozesmi, Tan & Ozesmi, 2006) when there are few observations per variable available in the data set. Artificial neural networks have been perceived as a black box (Lek & Guégan, 1999; but see Olden & Jackson, 2002b), can be computationally intensive and sensitive to the structure of the training data set (Ozesmi *et al*., 2006). It is also more difficult to determine the relationships and the strengths of the predictor variables with the response variable when using artificial neural networks relative to other statistical approaches (Brosse *et al*., 1999; Olden & Jackson, 2002b). However, methodologies can be used to identify the relationships between variables in artificial neural networks such as the Neural Interpretation Diagram (Ozesmi & Ozesmi, 1999), Garson's algorithm (Garson, 1991; Goh, 1995), Lek's algorithm (Lek *et al*., 1996), randomization tests (Olden & Jackson, 2002b) and partial derivatives (Gevrey *et al*., 2006).

The coefficient of determination (R^2) suggested that the performance of regression trees was very similar to that of multiple regression. Of all of the statistical approaches used, the RMSE of the models with the validation data sets was the highest for the regression tree. These high-RMSE values show that the models

generated by the regression trees were not good predictors of water temperature in independent data sets. Regression trees can perform well on complex, untransformed ecological data that consist of high-order interactions, multicollinearity and nonlinear relationships between predictor variables (De'ath & Fabricius, 2000; De'ath, 2002), and provide graphical interpretation of complex ecological interactions (De'ath & Fabricius, 2000). Some studies have documented the enhanced performance of regression trees over multiple regression (e.g. Rejwan *et al.*, 1999; De'ath & Fabricius, 2000). However, multiple regression will outperform regression trees in cases where a strong linear relationship exists between the variables (De'ath & Fabricius, 2000). Furthermore, only one mean predictive value is generated for all samples in the same leaf, such that all lakes in a specific leaf will have the same predicted water temperature. Depending on how the tree is pruned (e.g. minimum number of observations per leaf) and the size of the data set available, this may introduce greater amounts of error and reduces the strength of prediction.

Generally, based on the coefficient of determination, Bayesian multiple regression underperformed at predicting maximum near-surface water temperature. The Bayesian models selected tended to contain few variables, which may mean that they under-fitted the data. Furthermore, the Bayesian analyses were computationally time-consuming. The RMSE were not as large as found for the regression trees, however, suggesting that the models generated by Bayesian multiple regression were better predictors of maximum near-surface water temperature than the regression tree approach. It is also important to understand that Bayesian variable selection is currently a very active area of research (e.g. Kuo & Mallick, 1998; Casella & Moreno, 2006; Lunn *et al.*, 2006). Whereas the approach of Kuo & Mallick (1998) used here is relatively more complex than traditional multiple regression, it is one of the simpler Bayesian approaches to variable selection. As work continues, we can expect a greater understanding of variable selection which should lead to better statistical practice. Indeed, the fact that the very simple method of Kuo & Mallick (1998) performed at least moderately well suggests that more sophisticated Bayesian approaches hold promise for building statistical models to predict water temperatures.

Across the four modelling approaches used, artificial neural networks and stepwise multiple regression provided the best overall results, with the artificial neural networks providing the lowest RMSE in three of the four data sets. For the data set having the fewest observations and the greatest number of variables, the multiple-regression approach provided the best predictive capability, probably because of overfitting in the artificial neural network model in this one case. Across the four data sets these two modelling approaches provided comparable results. Given the simplicity of multiple regression, and the greater ease of interpreting the model terms (see Table 3), multiple regression can be recommended as a good method for similar data sets. However, it is the most sensitive of the four methods to departures from standard statistical assumptions (e.g. multi-collinearity and error distributions), so careful preparation of the data and the use of associated diagnostics remains essential. At present, we believe that traditional multiple regression, given its good predictive ability with our data sets and comparative simplicity, is preferable to Bayesian variable selection. The regression tree approach proved to have the greatest error rates associated with its predictions and cannot be recommended.

Ecological implications

We included climatic, geographic, lake morphology and water chemistry variables in our analyses to determine their relative importance in predicting lake temperatures. Previous studies had suggested (although those models were constructed with a small number of lakes restricted to a smaller geographical location) that variables such as mean depth (Shuter *et al.*, 1983; Snucins & Gunn, 2000; Edmundson & Mazumder, 2002), maximum depth (Kettle *et al.*, 2004), surface area (Kettle *et al.*, 2004), lake colour (Edmundson & Mazumder, 2002), turbidity (Edmundson & Mazumder, 2002) and dissolved organic carbon (Snucins & Gunn, 2000) played an important role in predicting maximum lake water temperatures, thus justifying the inclusion of lake morphology and water chemistry variables in our models. Our study suggested that, however, that among all statistical approaches and data sets, mean July air temperature and mean annual air

temperature, day of year and year were the most important predictor variables predicting maximum near-surface lake water temperatures. Lake morphology and water chemistry explained little variation in lake temperature.

We used the IPCC 1961–1990 average values for annual and July air temperature, which are summarized on a 0.5° × 0.5° grid. The use of regional 1961–1990 air temperature values incorporates a measure of space, in addition to climate, and permits the use of our predictive models with future Global Circulation Models (Sharma *et al.*, 2007). Air temperature has been linked empirically to surface-water temperature (e.g. Shuter *et al.*, 1983; Livingstone & Lotter, 1998; Livingstone & Dokulil, 2001; Livingstone & Padisák, 2007). We found that mean air temperature tended to be the most important variable in predicting maximum near-surface water temperature. In addition to our study, summer air temperature (i.e. July) has been linked to maximum surface water temperatures in lakes in the Swiss Plateau (Livingstone & Lotter, 1998) and Lake Superior (Austin & Colman, 2007). Mean annual air temperature is also an important predictor as an increase in the annual heat input into a lake should result in an increased maximum surface temperature (Shuter *et al.*, 1983). Our study did not reveal the importance of spring air temperature in predicting maximum near-surface water temperature, although others (e.g. Snucins & Gunn, 2000; Austin & Colman, 2007) have found that the highest summer water temperatures tended to be recorded in years that had relatively warmer springs. The rapid increase in spring heating may lead to higher water temperature (Snucins & Gunn, 2000) suggesting that intra-annual variability in air temperatures may play an important role in predicting lake temperature.

Day of year and year tended to be very strong predictors of maximum near-surface lake water temperature indicating that conditions related to the sampling period within the summer season or year were influential. Examination of water temperatures among years suggests that, since 1960, there has been a trend of increasing water temperature across Canadian lakes (Sharma *et al.*, 2007) which matches the general findings of others (e.g. French *et al.*, 2006; Austin & Colman, 2007). Ideally, air and water temperature would have been recorded simultaneously, but this was not possible. However, small-scale daily fluctuations in air temperature may have also simply provided additional noise in the analysis rather than picking up longer term signals. If air and water temperatures could have been measured concurrently, or appropriate air temperature time lags incorporated (*sensu* Matuszek & Shuter, 1996; Kettle *et al.*, 2004), air temperature may have been an even better predictor of lake temperature, thereby reducing the importance of day and year as predictor variables.

Across spatial scales, lake morphology played little role in predicting water temperature. Previous studies (e.g. Shuter *et al.*, 1983; Snucins & Gunn, 2000; Edmundson & Mazumder, 2002) incorporated lake morphology into their models, although these models explained little variation in near-surface water temperature. These models were generally developed on data sets with few observations and encompassed lakes spread over smaller areas than our study. We hypothesize that, at smaller spatial scales, lake morphology would be of greater importance in explaining maximum near-surface water temperatures than was found in our large-scale comparisons. Even our most geographically restricted data set encompassed lakes several hundred kilometres apart. Therefore, it appears that water temperature was dominated by large-scale-driving variables, such as climate or patterns in covariation of lake size with geography (e.g. the largest lakes tend to follow the diagonal southern boundary of the Canadian Shield). The relative importance of variables such as lake morphology and water chemistry tend to be emphasized at local scales where there is a more limited range in climate and radiation (*sensu* Jackson *et al.*, 2001).

The development of an effective predictive model to determine maximum lake near-surface water temperature is important for our understanding of lakes as ecological systems. In addition, water-temperature models can incorporate Global Circulation Models to predict the effects of climate change on future maximum lake temperature. Thus, Sharma *et al.* (2007) incorporated a variant of the multiple-regression model and air temperature predictions from the Canadian General Circulation Model 2 to predict maximum lake near-surface water temperature and the thermal habitat of smallmouth bass (*Micropterus dolomieu* Lacépède) in Canadian lakes in 2100.

Acknowledgments

We thank the organizations that contributed data to our study including: Ken Minns from the University of Toronto and Department of Fisheries and Oceans Canada, Brian Shuter from the University of Toronto and Ontario Ministry of Natural Resources, Keith Somers from the Ontario Ministry of Environment, Jenni McDermid from the University of Toronto, Jody McKenzie-Grieve and John Post at the University of Calgary, Dave Scruton and Mike de Jong from the Department of Fisheries and Oceans Canada, Ray Semkin and Fariborz Norouzian from the Canadian Centre of Inland Waters, John Gunn, Bill Keller, and Michael Malette from Laurentian University, Paul Blanchfield and researchers at the Experimental Lakes Area, Peter Leavitt at University of Regina, Beatrix Beisner and Maria Lorena Longhi at Université du Québec à Montreal, Daniel Boisclair and Pascale Gibeau at Université de Montreal, and the many additional individuals involved in field collections and database management. We thank the numerous scientists who published their data in scientific journals, theses or in online databases. We thank Dr Escobar for assistance with WINBUGS. We thank Drs Brian Shuter, Ken Minns, Ann Zimmerman, Harold Harvey, John Magnuson, Alan Hildrew and two anonymous reviewers for their comments on earlier versions of the manuscript. Funding for this research was provided by a Natural Sciences and Engineering Research Council of Canada Research Discovery Grant to DAJ and NSERC Scholarship to SW.

Supplementary material

The following supplementary material is available for this article:

Appendix S1. Bayesian multiple regression methodology.

This material is available as part of the online article from:http://www.blackwell-synergy.com/doi/abs/10.1111/j.1365-2427.2007.01881.x

(This link will take you to the article abstract).

Please note: Blackwell Publishing are not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

References

Anderson W.L., Robertson D.M. & Magnuson J.J. (1996) Evidence of recent warming and El Nino-related variations in ice breakup of Wisconsin lakes. *Limnology and Oceanography*, **41**, 815–821.

Arai T. (1981) Climatic and geomorphological influences on lake temperature. *Verhandlungen der Internationaler Vereinigung für Theoretische und Angewandte Limnologie*, **21**, 130–134.

Austin J.A. & Colman S.M. (2007) Lake Superior summer water temperatures are increasing more rapidly than air temperatures: a positive ice-albedo feedback. *Geophysical Research Letters*, **34**, L06604, doi:10.1029/2006GL029021.

Brandt S.B., Mason D.M., McCormick M.J., Lofgren B. & Hunter T.S. (2002) Climate change: implications for fish growth performance in the Great Lakes. *American Fisheries Society Symposium*, **32**, 61–76.

Brosse S., Guegan J.F., Tourenq J.N. & Lek S. (1999) The use of artificial neural networks to assess fish abundance and spatial occupancy in the littoral zone of a mesotrophic lake. *Ecological Modelling*, **120**, 299–311.

Casella G. & Moreno E. (2006) Objective Bayesian variable selection. *Journal of the American Statistical Association*, **101**, 157–167.

Casselman J.M. (2002) Effects of temperature, global extremes, and climate change on year-class production of warmwater, coolwater, and coldwater fishes in the Great Lakes basin. In: *Fisheries in a Changing Climate* (Ed. N.A. McGinn), pp. 39–60. Blackwell Publishing, Bethesda, Maryland.

Chen Y. & Jackson D.A. (1995) Robust estimation of mean and variance in fisheries. *Transactions of the American Fisheries Society*, **124**, 401–412.

Christie G.C. & Regier H.A. (1988) Measures of optimal thermal habitat and their relationship to yields for 4 commercial fish species. *Canadian Journal of Fisheries and Aquatic Sciences*, **45**, 301–314.

De'ath G. (2002) Multivariate regression trees: a new technique for modeling species-environment relationships. *Ecology*, **83**, 1105–1117.

De'ath G. & Fabricius K.E. (2000) Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, **81**, 3178–3192.

Edmundson J.A. & Mazumder A. (2002) Regional and hierarchical perspectives of thermal regimes in subarctic, Alaskan lakes. *Freshwater Biology*, **47**, 1–17.

French T.D., Campbell L.M., Jackson D.A., Casselman J.M., Scheider W.A. & Hayton A. (2006) Persistent organic pollutants in Lake Ontario Salmon: changes linked to source control, trophodynamics and a

warming climate. *Limnology and Oceanography*, **51**, 2794–2807.

Garson G.D. (1991) Interpreting neural-network connection weights. *Artificial Intelligence Expert*, **6**, 47–51.

Gevrey M., Dimopoulos I. & Lek S. (2006) Two-way interaction of input variables in the sensitivity analysis of neural network models. *Ecological Modelling*, **195**, 43–50.

Goh A.T.C. (1995) Back-propagation neural networks for modeling complex systems. *Artificial Intelligence of Engineering*, **9**, 143–151.

Guisan A. & Zimmermann N.E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147–186.

Jackson D.A., Peres-Neto P.R. & Olden J.D. (2001) What controls who is where in freshwater fish communities – the roles of biotic, abiotic, and spatial factors. *Canadian Journal of Fisheries and Aquatic Sciences*, **58**, 157–170.

Kettle H., Thompson R., Anderson N.J. & Livingstone D.M. (2004) Empirical modelling of summer lake surface temperatures in southwest Greenland. *Limnology and Oceanography*, **49**, 271–282.

Kuo L. & Mallick B. (1998) Variable selection for regression models. *Sankhya*, **60**, 65–81.

Lek S. & Guégan J.F. (1999) Artificial neural networks as a tool in ecological modeling, an introduction. *Ecological Modelling*, **120**, 65–73.

Lek S., Delacoste M., Baran P., Dimopolous I., Lauga J. & Aulagnier S. (1996) Application of neural networks to modelling nonlinear relationships in ecology. *Ecological Modelling*, **90**, 39–52.

Livingstone D.M. & Dokulil M.T. (2001) Eighty years of spatially coherent Austrian lake surface temperatures and their relationship to regional air temperature and the North Atlantic Oscillation. *Limnology and Oceanography*, **46**, 1220–1227.

Livingstone D.M. & Lotter A.F. (1998) The relationship between air and water temperatures in lakes of the Swiss Plateau: a case study with palaeolimnological implications. *Journal of Paleolimnology*, **19**, 181–198.

Livingstone D.M. & Padisák J. (2007) Large-scale coherence in the response of lake surface-water temperatures to synoptic-scale climate forcing during summer. *Limnology and Oceanography*, **52**, 896–902.

Lunn D.J., Whittaker J.C. & Best N. (2006) A Bayesian toolkit for genetic association studies. *Genetic Epidemiology*, **30**, 231–247.

Magnuson J.J., Crowder L.B. & Medvick P.A. (1979) Temperature as an ecological resource. *American Zoologist*, **19**, 331–343.

Magnuson J.J., Meisner J.D. & Hill D.K. (1990) Potential changes in the thermal habitat of Great Lakes fish after global climate warming. *Transactions of the American Fisheries Society*, **119**, 254–264.

Matuszek J.E. & Shuter B.J. (1996) An empirical model for the prediction of daily water temperatures in the littoral zone of temperate lakes. *Transactions of the American Fisheries Society*, **125**, 622–627.

McCombie A.M. (1959) Some relations between air temperatures and the surface water temperatures of lakes. *Limnology and Oceanography*, **4**, 252–258.

Ntzoufras I.P. (2002) Gibbs variable selection using BuGS. *Journal of Statistical Software*, **7**, 1–19.

Olden J.D. & Jackson D.A. (2000) Torturing the data for the sake of generality: how valid are our regression models? *Ecoscience*, **7**, 501–510.

Olden J.D. & Jackson D.A. (2001) Fish-habitat relationships in lakes: gaining predictive and explanatory insight by using artificial neural networks. *Transactions of the American Fisheries Society*, **130**, 878–897.

Olden J.D. & Jackson D.A. (2002a) A comparison of statistical approaches for modelling fish species distributions. *Freshwater Biology*, **10**, 1976–1995.

Olden J.D. & Jackson D.A. (2002b) Illuminating the ''black box'': understanding variable contributions in artificial neural networks. *Ecological Modelling*, **154**, 135–150.

Olden J.D., Joy M.K. & Death R.G. (2006) Rediscovering the species in community-wide predictive modeling. *Ecological Applications*, **16**, 1449–1460.

Ozesmi S.L. & Ozesmi U. (1999) An artificial neural network approach to spatial habitat modeling with interspecific interaction. *Ecological Modelling*, **116**, 15–31.

Ozesmi S.L., Tan C.O. & Ozesmi U. (2006) Methodological issues in building, training, and testing artificial neural networks in ecological applications. *Ecological Modelling*, **195**, 83–93.

Pearce J. & Ferrier S. (2000) An evaluation of alternative algorithms for fitting species distribution models using logistic regression. *Ecological Modelling*, **128**, 127–147.

Rejwan C., Collins N.C., Brunner L.J., Shuter B.J. & Ridgway M.S. (1999) Tree regression analysis on the nesting habitat of smallmouth bass. *Ecology*, **80**, 341–348.

Sharma S. & Jackson D.A. (2008) Predicting smallmouth bass incidence across North America under climate change: a comparison of statistical approaches. *Canadian Journal of Fisheries and Aquatic Sciences*, in press.

Sharma S., Jackson D.A., Minns C.K. & Shuter B.J. (2007) Will northern fish populations be in hot water because of climate change? *Global Change Biology*, **13**, 2052–2064, doi:10.1111/j.1365-2486.2007.01426.x.

Shuter B.J. & Post J.R. (1990) Climate, population viability, and the zoogeography of temperate fishes. *Transactions of the American Fisheries Society*, **119**, 314–336.

Shuter B.J., MacLean J.A., Fry F.E.J. & Regier H.A. (1980) Stochastic simulation of temperature effects on first-year survival of smallmouth bass. *Transactions of the American Fisheries Society*, **109**, 1–34.

Shuter B.J., Schlesinger D.A. & Zimmerman A.P. (1983) Empirical predictors of annual surface water temperature cycles in North American lakes. *Canadian Journal of Fisheries and Aquatic Sciences*, **40**, 1838–1845.

Snucins E. & Gunn J. (2000) Interannual variation in the thermal structure of clear and colored lakes. *Limnology and Oceanography*, **45**, 1639–1646.

Staehr P.A. & Sand-Jensen K. (2006) Seasonal changes in temperature and nutrient control of photosynthesis, respiration and growth of natural phytoplankton communities. *Freshwater Biology*, **51**, 249–262.

Tonn W.M. (1990) Climate change and fish communities: a conceptual framework. *Transactions of the American Fisheries Society* **119**, 337–352.

Uimari P. & Hoeschele I. (1997) Mapping linked quantitative trait loci using Bayesian method analysis and Markov chain Monte Carlo algorithms. *Genetics*, **146**, 735–743.

Whittington M.J., Stephens P.A., Bradbury R.B. & Freckleton R.P. (2006) Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology*, **75**, 1182–1189.

(Manuscript accepted 1 December 2007)