J. Great Lakes Res. 27(4):486–502
Internat. Assoc. Great Lakes Res., 2001

Tree-Structured Modeling of the Relationship Between Great Lakes Ice Cover and Atmospheric Circulation Patterns

Sergei Rodionov, Raymond Assel,* and Lynn Herche

Great Lakes Environmental Research Laboratory,
2205 Commonwealth Blvd.
Ann Arbor, Michigan 48105

ABSTRACT. *Seasonal maximum ice concentration (percentage of lake surface covered by ice) for the entire Laurentian Great Lakes and for each Great Lake separately is modeled using atmospheric teleconnection indices. Two methods, Linear Regression (LR) and Classification and Regression Trees (CART), are used to develop empirical models of the interannual variations of maximum ice cover. Thirty-four winter seasons between 1963 and 1998 and nine teleconnection indices were used in the analysis. The ice cover characteristics were different for each Great Lake. The ice cover data lent itself better to CART analysis, because CART does not require* a priori *assumptions about data distributions characteristics to perform well. The stepwise LR models needed more variables, and in general, did not explain as much of the variance as the CART models. Two variables, the Multivariate ENSO index and Tropical/Northern Hemisphere index, explained much of the interannual variations in ice cover in the CART models. Composite atmospheric circulation patterns for threshold values of these two indices were found to be associated with above-and below-normal ice cover in the Great Lakes. Thus, CART also provided insight into physical mechanisms (atmospheric circulation characteristics) underlying the statistical relationships identified in the models.*

INDEX WORDS: *Ice cover, Great Lakes, teleconnections, ice models, atmospheric circulation.*

INTRODUCTION

The Laurentian Great Lakes (Fig. 1) are located in the mid-latitudes of North America. The ice cover that forms during the winter affects the ecology and economy of the region. Ice cover impacts the winter lake aquatic system (Magnuson *et al.* 1997), winter shipping activity (Assel *et al.* 2000), hydropower generation (International Niagara Working Committee 1983), shore installations (Wortley 1978), lake effect snowfall (Burrows 1991), and lake evaporation and water levels (Croley and Assel 1994).

Atmospheric circulation is a key factor affecting the temporal and spatial variability of ice cover through mass and energy transport across the lake—atmosphere boundary. The mechanisms of this control are both direct and indirect. A direct, or mechanical, effect of wind on ice cover includes the destruction of the ice field and spatial redistribution of ice (e.g., piling it up on one or another side of the lake). One strong storm in the beginning of ice season can completely destroy a newly formed ice cover even if air temperature is below freezing. In addition to the direct impact, strong winds increase vertical mixing of water thereby increasing water temperature in the surface layer.

Frequency of storms over a lake and the direction of air advection are related to large-scale atmospheric circulation patterns. Rohli *et al.* (1999) have examined the linkage between regional scale atmospheric circulation in the Great Lakes basin and continental- to hemispheric-scale circulation patterns. Their results support the notion that the regional atmosphere undergoes shifts consistent with the broader-scale circulation. A consideration of atmospheric circulation on a *space scale* larger than the lake basin becomes particularly important as the *time scale* of regional processes under investigation increases. In this study the characteristics of ice cover that pertain to the entire winter season are used to study variations of ice cover from one season to another. The interannual variations in ice cover are a climatological problem and require a global perspective.

*Corresponding author: E-mail: Assel@glerl.noaa.gov

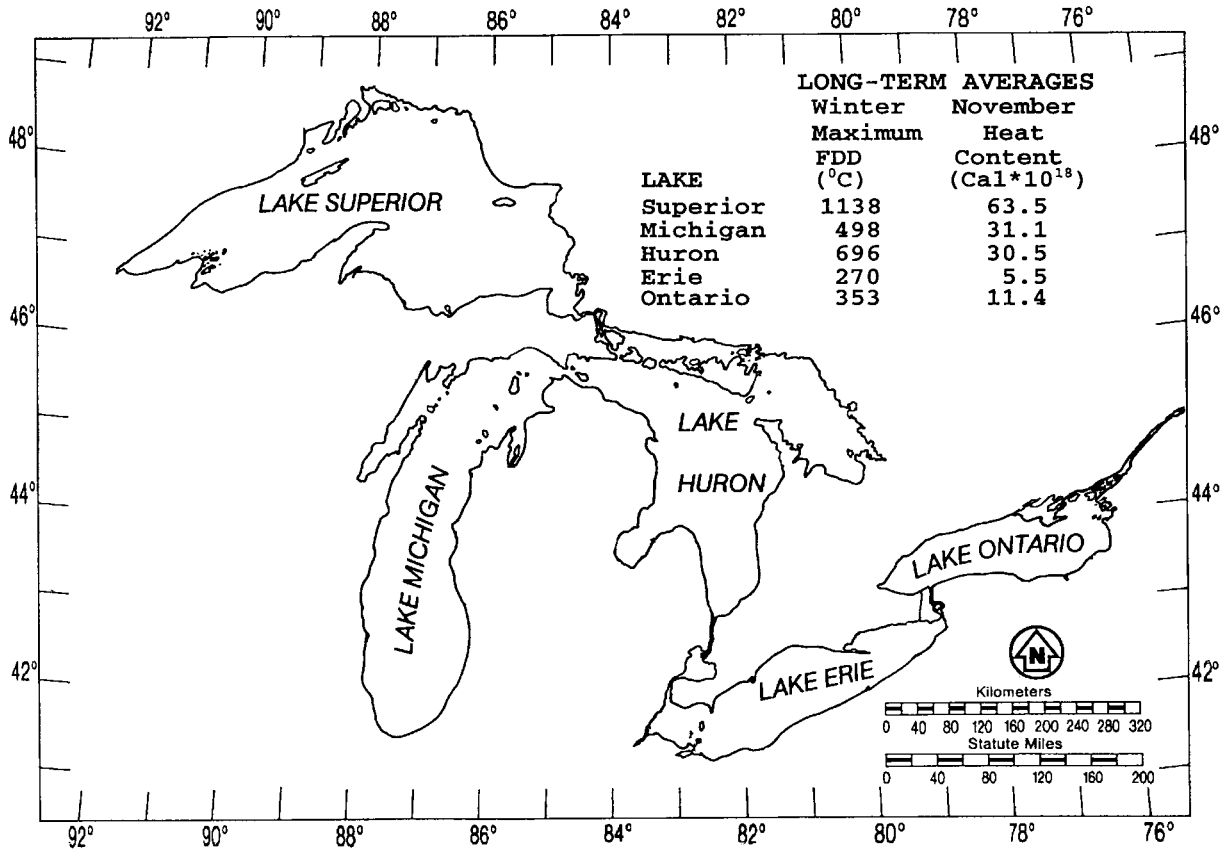

FIG. 1. Lake averages (winters 1963 to 1998) of maximum Freezing Degree-Day (FDD) and November heat storage. The FDD accumulations on each Great Lake are calculated from data given in Assel* et al. *(2000) and lake averages (1962 to 1995) of November heat content is calculated from Croley (1992).

The interannual variability in the wintertime atmospheric circulation is dominated by a number of modes that stand out above the background continuum (Kushnir and Wallace 1989). These modes are also known as teleconnections because they exhibit in-phase or out-of-phase correlations between regions (often called "centers of action") separated by large distances. Each teleconnection pattern is characterized by an index that is most commonly calculated as a combination of 500/700-hPa geopotential height anomalies in the centers of action or as principal component scores of the corresponding empirical orthogonal function.

From a practical point of view teleconnection indices are a compact, parametric way to describe the complex dynamics of large-scale atmospheric circulation. They are particularly useful in modeling the relationships with regional climatic characteristics. For example, Yarnal and Leathers (1988) have found that the interannual variability of the Pennsylvania climate is related to two important Northern Hemisphere teleconnections: the Pacific/North American (PNA) and North Atlantic Oscillation (NAO) patterns. Hartley and Keables (1998) have demonstrated that winter snowfall in New England is associated with a meridional circulation regime, as indicated by a negative NAO index. They have found no direct association with the PNA index, but noticed that it can modulate the association with the NAO. According to Rohli *et al.* (1999), the regional atmosphere over the Great Lakes experiences the effect of both the PNA and NAO patterns. Evidence of PNA teleconnections with Great Lakes climatic variables was presented by Assel (1992). In a previous paper Assel and Rodionov (1998) found that Great Lakes ice cover is most strongly correlated with the Tropical/Northern Hemisphere (TNH)

index, particularly for the detrended time series on Lake Michigan.

The El Niño/Southern Oscillation (ENSO) impacts on the North American climate stimulated research on implications for the Great Lakes. The 1982/83 El Niño event (Assel *et al.* 1985) had much below average ice cover in the lakes as did the 1997/98 El Niño event (Assel 1998). Records for winters 1963 through 1990 (Assel and Rodionov 1998) showed that 46% of the lowest quartile of annual maximum ice covers occurred during El Niño events. The atmospheric flow over the Great Lakes appears to be less disturbed during El Niño events (Angel *et al.*1999) with less frequent cyclones during El Niño years than non-ENSO years. Rohli *et al.* (1999), however, have found no strong association between Great Lakes regional surface circulation and ENSO events. This may be a result of a non-linear response of Great Lakes climatic variables to ENSO events.

The purpose of this study is to compare empirically two methods of modeling the relationship between ice cover and atmospheric teleconnection patterns. The first method is a conventional linear regression analysis. The linear regression (LR) model is usually used as the first choice if no *a priori* information exists about the functional form of the relationship. The second method is called Classification and Regression Tree (CART) analysis. As the name suggests, CART is a single procedure that can be used to analyze either categorical (classification) or continuous (regression) data. Classification trees were used (Rodionov and Assel 1999) to develop an empirical classification of atmospheric circulation patterns associated with below-normal, normal, and above-normal ice cover in the Great Lakes. Here CART will be used to analyze a relationship between atmospheric circulation and ice cover presented as continuous variables in regression trees. CART and linear regression models will be compared in terms of their accuracy, complexity, and ability to provide meaningful interpretation of the results and to get insight into the mechanisms of the relationship.

DATA

Ice Cover

The ice cover data used in this study represent a set of 34 values of maximum winter ice cover for winters 1963 through 1998, excluding winters 1996 and 1997, for which data were not available at the time of study. Maximum winter ice cover was used in this study because of its relatively long period of record of large-scale ice extent on the Great Lakes. Estimates of maximum ice extent were made from operational ice charts produced by the National Ice Center, U.S. Coast Guard, U.S. Army Corps of Engineers, National Oceanic and Atmospheric Administration, and the Canadian Ice Service (Assel and Rodionov 1998). The total ice coverage of the five Great Lakes (expressed as a percent of total surface area covered by ice) is calculated as the weighted sum of the ice-covered areas of the five Great Lakes.

Ice cover formation can be considered to be a threshold process (Assel 1991). Ice cover extent remains low until the number of freezing degree-days (FDDs) reaches a certain threshold, after which ice cover experiences rapid growth. The threshold is basically a function of geographical position and heat storage of the lake. Thus, for Lake Superior, the northernmost among the Great Lakes (Fig. 1), seasonal accumulations of FDDs usually exceeds the threshold, and annual maximum ice cover is frequently greater than 70% (Fig. 2b). As a result, the median value of maximum ice cover on Lake Superior is higher than the mean value, and the skewness is negative (Table 1). The distribution of annual maximum ice cover for Lake Erie (Fig. 2e) is even more skewed to the left with only one out of every four winters having less than 90% ice cover. Despite its southernmost position and the highest surface air temperature, it is often almost entirely covered by ice. Unlike Lake Superior, however, the major factor driving negative skewness is the relatively low heat storage of this shallow lake. In contrast to Lakes Superior and Erie, maximum ice covers for Lakes Michigan and Ontario (Fig. 2c and 2f) are right skewed. Lake Ontario with much larger heat storage than Lake Erie but only marginally higher freezing degree-days (Fig. 1) has the lowest maximum ice cover among the Great Lakes with half of the winters not exceeding 20%. Lake Michigan is significantly elongated from north to south with ice mostly in the north and usually less than 50%. Lake Huron (Fig. 2d) is the only lake with the median close to the mean. It features a strongly negative kurtosis (Table 1) and is close to a uniform distribution. Considering the Great Lakes together and combining ice cover proportionally to surface area, the ice cover distribution is similar to Lake Huron (Fig. 2a).

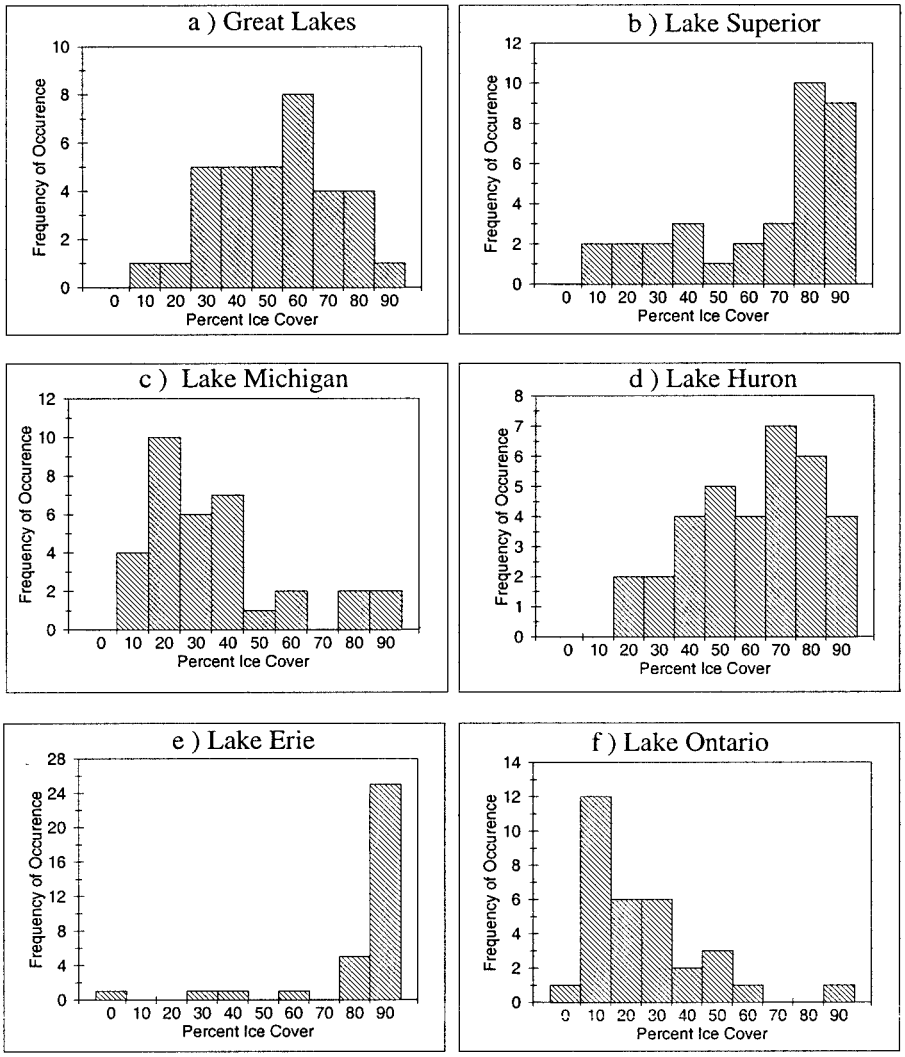

FIG. 2. Histograms and empirical frequency distribution functions for ice cover in the Great Lakes.

Teleconnection Indices

Teleconnection indices used as independent variables are as follows:

1) The Polar/Eurasian index (POL)
2) The West Pacific index (WP)
3) The East Pacific index (EP)
4) The Pacific/North American index (PNA)
5) The Tropical/Northern Hemisphere index (TNH)
6) The North Atlantic Oscillation index (NAO)
7) The East Atlantic index (EA)
8) The Multivariate ENSO Index (MEI)
9) The Southern Oscillation Index (SOI)

The first seven of these indices are regularly calculated by the Climate Prediction Center (CPC), of NOAA and are available over the Internet[1]. In this study mean winter (DJF) values were used of all the indices except the TNH, for which February data were absent, and so the mean value of December and January was used. The diagnostic procedure used by the CPC to identify teleconnection patterns is the Rotated Principal Component analysis—RPC (Barnston and Livezey 1987). The teleconnection patterns are identified based on the entire flow field, not just from height anomalies at a few select

[1] http://www.cpc.noaa.gov/data/indices/

TABLE 1. Statistical characteristics of ice cover variations in the Great Lakes.

Statistic	All Lakes	Superior	Michigan	Huron	Erie	Ontario
Mean	58	69	39	66	87	27
Median	59	81	32	68	95	20
Std. Dev.*	20	27	23	22	21	20
Skewness	0	–0.9	1.2	–0.1	–2.8	1.5
Kurtosis	–0.4	–0.6	0.8	–1.2	7.9	2.9

* Standard Deviation—rounded to the nearest whole number

locations. The VARIMAX rotation procedure used by the CPC retains the temporal orthogonality between the RPCs (Kushnir and Wallace 1989). Therefore, teleconnections can be considered as *independent* modes of large-scale atmospheric circulation.

To characterize ENSO events two indices were used: 1) the Southern Oscillation Index (SOI), and 2) the Multivariate ENSO Index (MEI). The SOI was used in its standard form as difference in SLP between Tahiti and Darwin. These data are also available from the CPC web site but they are not independent of each other and no claim is made for their independence of the seven indices noted above. The MEI (courtesy of K. Wolter) can be understood as a weighted average of the main ENSO features contained in the following six variables: sea-level pressure, the east-west and north-south components of the surface wind, surface sea tempearature, surface air temperature, and total amount of cloudiness. Positive (negative) values of the MEI represent the warm (cold) ENSO phase. The MEI is computed separately for each of 12 sliding bimonthly seasons (Dec/Jan, Jan/Feb, . . . , Nov/Dec). All seasonal values are standardized with respect to both season and the 1950 to 1993 reference period. In this study only Dec/Jan values of the MEI were used (Wolter and Timlin 1998).

CART METHOD

CART is an analysis tool for partitioning data (Breiman *et al.* 1984). Some recent applications in atmospheric and hydrological sciences include: Burrows and Assel 1992, Rodionov 1994, Spear *et al.* 1994, Burrows *et al.* 1995, Zorita *et al.* 1995. CART presents its results in the form of decision trees using methodology known as binary recursive partitioning. The process is binary because parent nodes are always split into exactly two child nodes, and recursive in repeatedly treating each child node as a potential parent.

CART's goal in forming a regression tree is to partition the data into homogeneous (low variance) terminal nodes; the mean value in each node is its predicted value. The tree-growing process is based on the least-squared deviation method of finding the best split at each node using the "improvement" (or reduction) in variance due to the split, calculated as:

$$\text{Improvement} = \sigma^2_p - (n_l/n_p \times \sigma_l^2 + n_r/n_p \times \sigma_r^2),$$

where σ^2_p, σ_l^2, and σ_r^2 (n_p, n_l, and n_r) are variances (number of observations) respectively, in the parent, left child, and right child nodes. The best split *s* maximizes the improvement.

CART and LR models have an important common feature. CART fits a simple 1 factor two level Analysis of Variance (ANOVA) model to a group of observations as it computes the variance within the observation partition based on a split of the independent variable X. The values of the (final) group means are used in post-hoc predictions. A simple regression extends such an ANOVA to a proportional relation (plus constant) between the X and Y variables instead of "dummy" variables, which merely encode the partition. CART and LR differ, however, in two fundamental respects. First, unlike ANOVA in which the partition is fixed in advance, CART defines the groups by *using* the variance minimization (or, equivalently, maximizing the "goodness of fit") thus resulting in a very nonlinear relation. Second, unlike ANOVA which fits one relation to the whole set of observations, CART only does this for single nodes, which are, except for the root node, proper subsets of the whole.

One crucial question in CART analysis is when to stop growing a tree, i.e., when to stop splitting nodes. As in most other statistical analyses it is a trade-off between predictive precision and model simplicity (Bohanec and Baratko 1994). With regression trees, simplicity is obtained by selecting trees with fewer terminal nodes, while precision is

gained by allowing more terminal nodes. Various stopping rules can be used to automate CART, such as minimum number of cases in a node, minimum improvement value, or maximum depth of the tree. Sometimes these stopping rules produce poor results, because a node that might not split well at one level might yield very informative splits if the tree-growing process continued just a little further. Therefore, it is often recommended to grow a maximal tree when further splitting becomes impossible and then prune it back. The best tree is the one that strikes a balance between predictive precision and comprehensibility. From ice cover data available for this analysis (34 winters), the final trees were not more than three levels deep and the number of independent variables in the CART models was equal or less than that in the corresponding LR models.

CART uses two methods to assess its accuracy. The preferred method is to use a separate test data set. When data are scarce, CART uses cross-validation to assess its goodness of fit. The learning sample is divided into 10 roughly equal parts. CART takes the first nine parts of the data (the learning sample), constructs the largest possible tree, and uses the remaining 1/10 of the data to obtain initial number of misclassifications (error rate) of selected sub-trees. The same process is then repeated on another 9/10 of the data while using a different 1/10 part as the test sample. The process continues until each part of the data has been held in reserve one time as a test sample. The results of the 10 mini-test samples are then combined to form error rates for trees of each possible size; these error rates are applied to the tree based on the entire learning sample. In the CART software used, cross-validation could only be accomplished when CART automatically grows the tree. Trees were not grown automatically to permit more control over the modeling process.

LINEAR REGRESSION

The models made by the multiple linear regression for comparison were composed by the forward stepwise method from the complete list of independent variables. The selection list was terminated at the variable for which R-squared was not essentially increased by using an additional variable. The choice was deliberately made to liberally include variables to reduce any chance that comparison with CART would be influenced by missing a "critical" variable. The choice of a model was not significantly dependent on the choice of forward, backward, or true stepwise selection nor on the entry or exit criteria, given the liberal selection policy.

The ice cover data was transformed (Arc Sin of Square Root of [% ice cover/100]) prior to the LR analysis to insure the LR models would not predict ice cover greater than 100% or less then 0%. The regression coefficients given below are for the transformed data. The modeled ice cover was transformed back to percent ice cover for comparison with CART modeled ice cover.

RESULTS

The CART models for each of the Great Lakes and for combined ice cover for all the lakes are presented in Figure 3. The characteristics of these models are summarized in Table 2. The same characteristics for the LR models are also presented. The CART models are often more compact than the corresponding LR models in terms of the number of parameters. When constructing the LR models, the complexity of the model was limited by an empirical rule that a variable to be added to the model should increase the percentage of explained variance by at least 4%. As the result, the final LR models have 3 to 4 parameters. The CART models have 2 or 3 parameters, but demonstrate equal or better performance than the LR models (Table 2).

Lake Superior

In the CART model for Lake Superior (Fig. 3a) the first split on variable MEI separates a group of nine winters for which MEI is greater than 0.8, i.e., winters of strong El Niño events. It was observed that even a simple model with just one split can describe a significant portion of variation (Fig. 4a). As previously noted, the number of FDDs for this lake usually exceeds the threshold for extensive ice cover. As a result, maximum ice cover usually fluctuates around its median value of 80% with occasional "spikes" of very mild winters. Six winters had ice cover below 35% (1964, 1975, 1983, 1987, 1995, and 1998); all but 1975 were winters of strong El Niño events. Figure 4a shows that a simple tree-structured model with just two output values (one for strong El Niño events and one for the rest) can be a good first approximation of ice cover on Lake Superior describing about 44% of the total variance. Note that the MEI index for the LR model

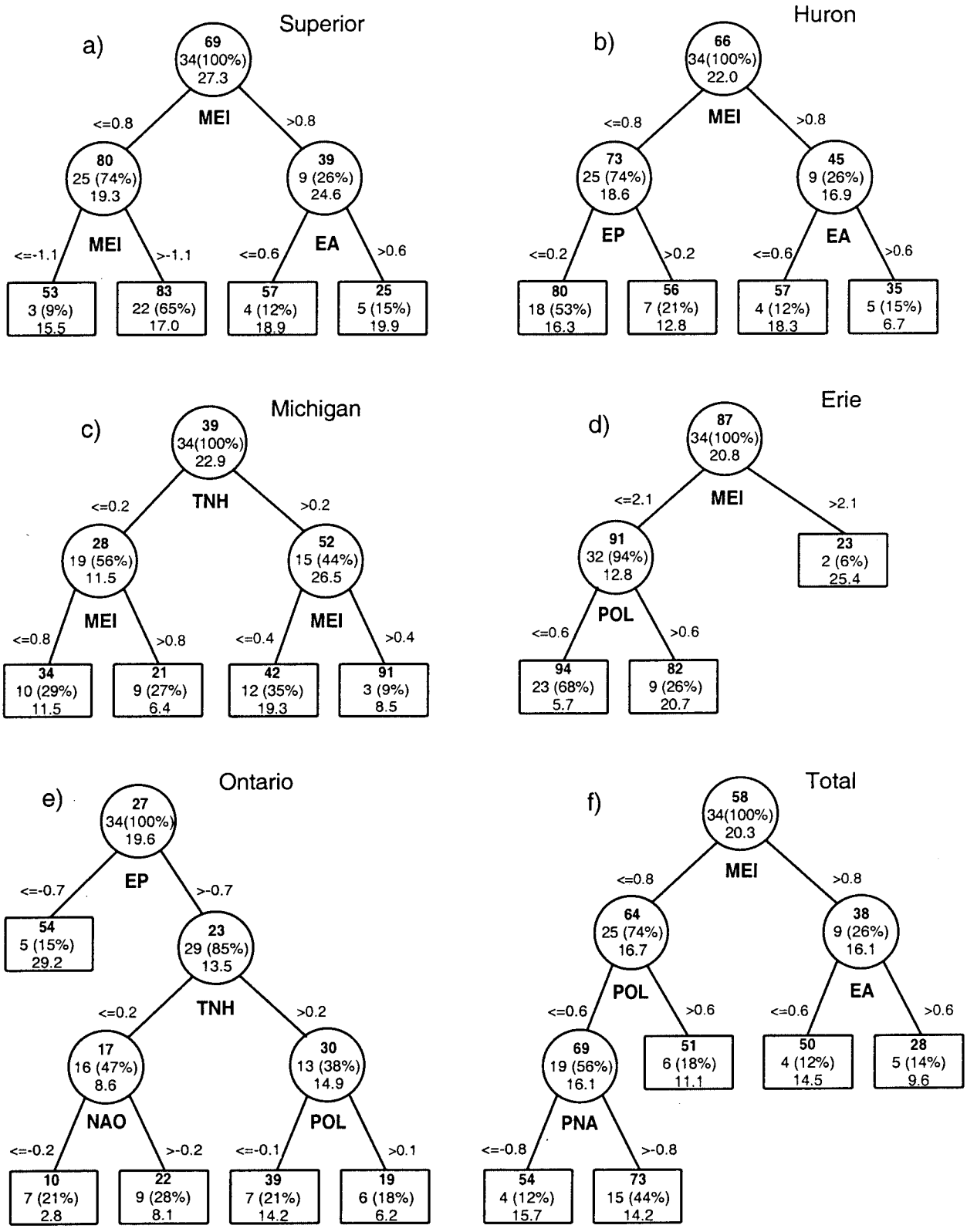

FIG. 3. ***CART models for ice cover in the Great Lakes. Numbers in each node are average value, number of cases, percent of cases (in parentheses), and standard deviation.***

TABLE 2. Model parameters and performance characteristics: percent of explained variance (R^2) and standard error. For linear regression models numbers in parentheses show the percentage of explained variance given that preceding variables are in the model.

Lake	Model	Independent variables	R2	Standard error
Superior	CART	MEI, EA	63	16.4
	LR	TNH(29), EP(14), POL(17), MEI(6)	66	16.0
Huron	CART	MEI, EP, EA	58	14.1
	LR	TNH(20), EP(19), POL(6), SOI(4)	49	15.5
Michigan	CART	TNH	29	18.0
	CART	TNH, MEI	66	13.0
	LR	TNH(24), POL(12), EP(14), PNA(5)	55	14.4
Erie	CART	MEI, POL	67	11.9
	LR	MEI(19), EP(8), POL(7)	34	16.6
Ontario	CART	EP, TNH	43	14.5
	CART	EP, TNH, POL	53	13.2
	CART	EP, TNH, POL, NAO	58	12.6
	LR	EP(21), TNH(15), POL(10)	49	13.7
All Lakes	CART	MEI, POL, EA, PNA	62	12.3
	LR	TNH(22), EP(16), POL(14), SOI(5)	57	13.5

(ICE_SUP = – 0.13MEI + 1.04) accounts for only 19% of total variance (Fig. 4b).

Figure 4b reveals the deficiency of a simple regression model that links Lake Superior ice cover with the MEI index. This is a typical problem that occurs with LR modeling. Since the model tries to find the best fit based on all the data including the outliers, it results in a systematic under-valuation of most of the data and in an overvaluation of the extremely mild winters. Adding EA to the LR model improves the percentage of explained variance to 43%, almost doubling the explained variance and comparable to the simplest CART model with two terminal nodes. This is, however, much lower than 63% for the CART model with the same two variables (Fig. 4c). It takes two more parameters, POL and TNH, for the LR model to reach the accuracy of the simpler CART model (Table 2 and Fig. 4d). The best LR model (Fig. 4d) with four independent variables is: ICE_SUP = 0.15TNH – 0.25EP – 0.17POL – 0.09MEI + 1.03. The LR model failed by more than 30% for five winters (1967, 1969, 1976, 1987, and 1992).

Note that for the CART model in Figure 3a the majority of cases (65%) are grouped in one relatively compact node with mean value close to the overall median. Three other terminal nodes contain anomalous winters. Figure 4c demonstrates the improvement of modeling the anomalous winters compared to the model with just two terminal nodes (Fig. 4a). The percentage of the total variance increased from 44% to 63%. Particular improvement was achieved for extremely mild winters of 1983, 1987, 1995, and 1998. Three winters (1966, 1969, and 1975) had absolute error of more than 30%.

Lake Huron

Ice cover on Lake Huron turned out to be difficult to model with both methods. The CART model for Lake Huron (Fig. 3b) is similar to that for Lake Superior (Fig. 3a) with two of three variables being the same. The coldest 18 winters (53%) in one node had average ice cover of 80%. Three mild winters (1968, 1969, and 1991) also included should not belong to this group with ice cover of only 50%. A failed attempt to separate these on the next step suggests they have little in common in terms of atmospheric circulation modes. Despite this relatively poor performance compared to other lakes, it still significantly outperforms the LR model for Lake Huron with four independent variables: ICE_HUR = 0.11TNH – 0.24EP – 0.09POL + 0.08SOI + 0.99. This LR model is very similar to that for Lake Superior, except that the MEI index is replaced by another ENSO index – SOI (Fig. 5b). In spite of the similarity of the LR models for Lakes Superior and Huron, the latter describes only 49% of total variance in ice cover compared to 66% for the former. Two years (1967 and 1991) had absolute error of more than 30%.

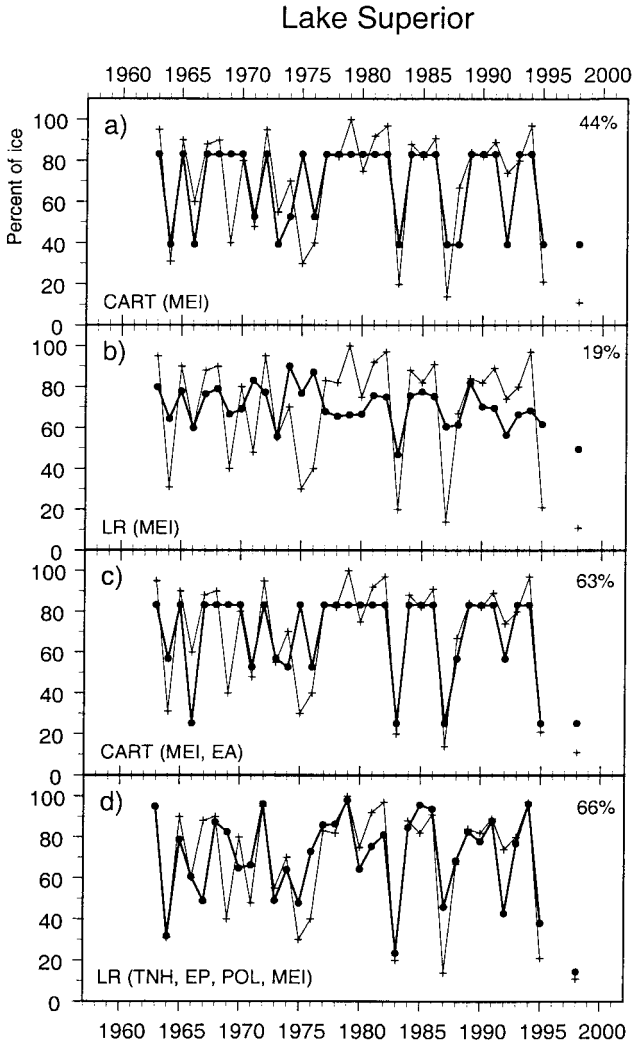

FIG. 4. Observed (+) and modeled (●) ice cover in Lake Superior. Numbers in top right corner are percent of explained variance. a) CART model with one split on MEI, b) LR model with MEI, c) CART model from Figure 3a, d) LR model with TNH, EP, POL, and MEI.

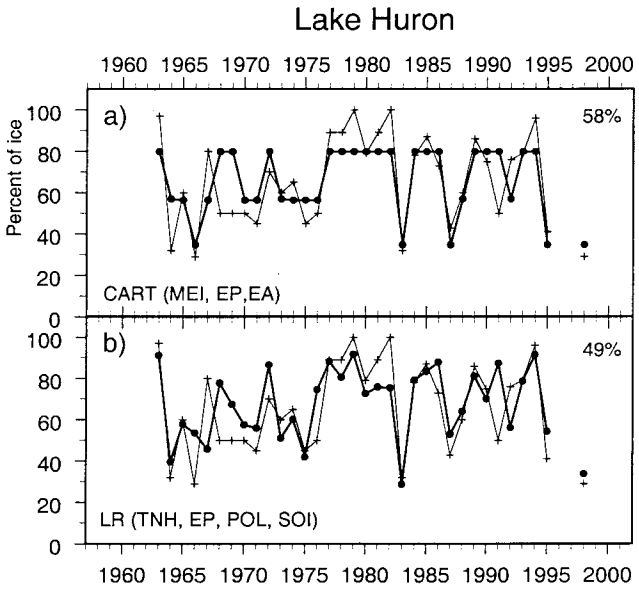

FIG. 5. Same as Figure 4, except for Lake Huron. a) CART model from Figure 3b, b) LR model with TNH, EP, POL, and SOI.

Lake Michigan

The CART model for Lake Michigan contains two independent variables, TNH and MEI (Fig. 3c). TNH was chosen first by both CART and LR providing an interesting comparison of simplified models for both methods (Table 2, Fig. 6a, Fig. 6b). The CART model (Fig. 6a) splits on TNH giving a mild winter value of 28% of ice cover or a cold winter value of 52%. Note the approximately equal numbers of mild and cold winters identified by the split, in contrast with models for Lakes Superior and Huron. Variability is much higher for cold winters than for mild winters. The model describes 29% of the total variance. The simple LR model (Fig. 6b) (ICE_MIC = 0.14TNH +0.68) describes 24% of the total variance.

The CART model improves significantly with one more level. The percentage of the explained variance with four terminal nodes (Fig. 3c) increases to 66%. Particular improvement (Fig. 6c) was for the three coldest years on the lake (1977, 1979, and 1994) in one separate node. The fourth coldest winter (1963) was not in this group, the only winter when the model failed by more than 30%. Note that for the winter of 1978, ice cover on Lake Michigan was less than for the other Great Lakes, and the CART model reflects this.

The best LR model (Fig. 6d, ICE_MIC = 0.17TNH – 0.17POL – 0.16EP + 0.09PNA + 0.67) is less accurate than the two variable CART model describing only 55% of the total variance (11% less than the CART model). The LR model failed on two winters (1979) (1991) by more than 30%.

Lake Erie

Lake Erie is another lake (with Lakes Superior and Huron), where CART first uses MEI (Fig. 3d). The split point, however, is different, and splits off

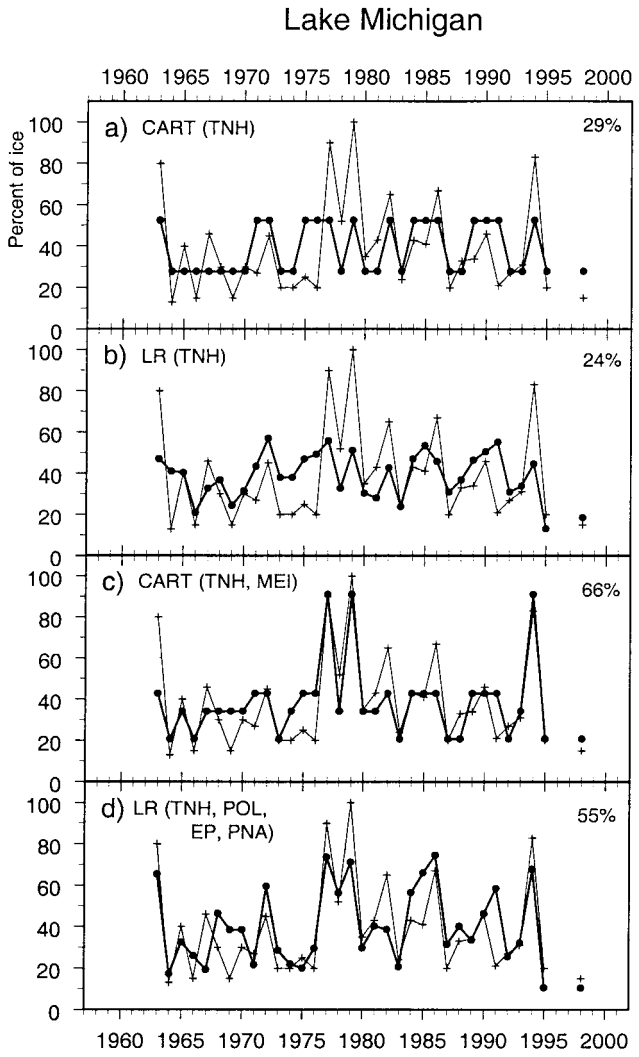

FIG. 6. Same as Figure 4, except for Lake Michigan. a) CART model with one split on TNH, b) LR model with TNH, c) CART model from Figure 3c, d) LR model with TNH, POL, EP and PNA.

a group of only two winters (1983, 1998) when the two strongest El Niño events of the century occurred. The second split (left) was made on POL, separating a group 23 (or 68%) winters with the average ice cover of 94%. The remaining nine mild winters have rather high variability. Overall, this simple model (Fig. 7a) with just two splits describes 67% of the total variance.

Lake Erie is the only lake where MEI was chosen by the LR model: ICE_ERI = –0.13MEI – 0.14EP – 0.10POL + 1.28. The MEI itself describes 19% of the total variance; two other variables, EP

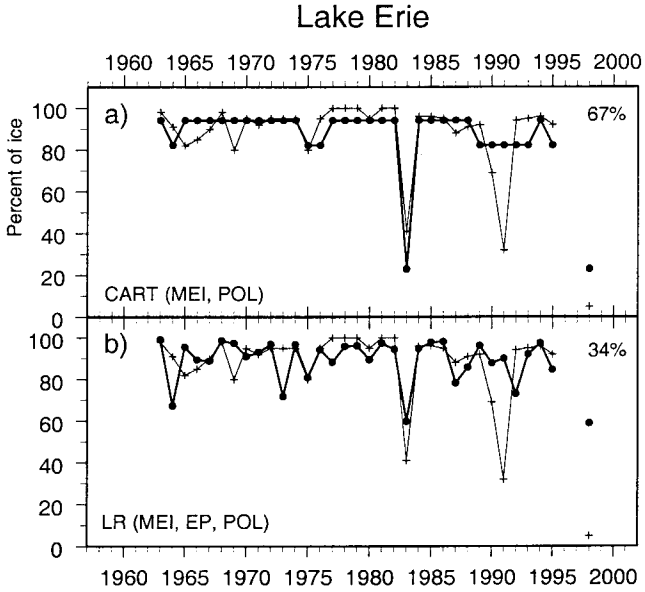

FIG. 7. Same as Figure 4, except for Lake Erie. a) CART model from Figure 3d, b) LR model with MEI, EP, and POL.

and POL, increase that to 34%. This is still about half that achieved by CART. Figure 7b shows that the LR model substantially over predicted in 1991 and 1998.

Lake Ontario

Ice conditions on Lake Ontario are much less severe than on the other Great Lakes with maximum annual ice cover around 25% and rarely reaching 50%. The lake was almost completely frozen only once (1979) (Fig. 8a). CART first (Fig. 3e) separates cold winters using EP. This was also the first parameter chosen for the LR model: ICE_ONT = –0.18EP + 0.11TNH –0.10POL + 0.53, which captures 49% of the total variance (Table 2, Fig. 8b). A CART model of two splits (EP, TNH) has 43% of the total variance (Table 2). The node with highest mean ice (54%) is the chief impediment to improving the model. It had the 3 coldest years (1978, 1979, 1994) and 2 years (1993, 1995) with medium and low cover. Apparently minor shifts in atmospheric flow patterns usually associated with cold winters (meridional circulation—Assel and Rodionov 1998) can result in average or mild winters, making it difficult to separate cold winters into one node.

When CART uses the same variables used by LR

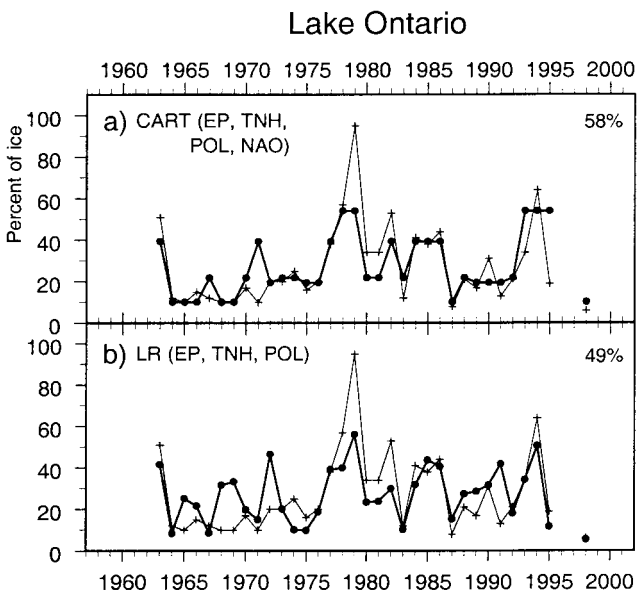

FIG. 8. Same as Figure 4, except for Lake Ontario. a) CART model from Figure 3e, b) LR model with EP, TNH, POL, and NAO.

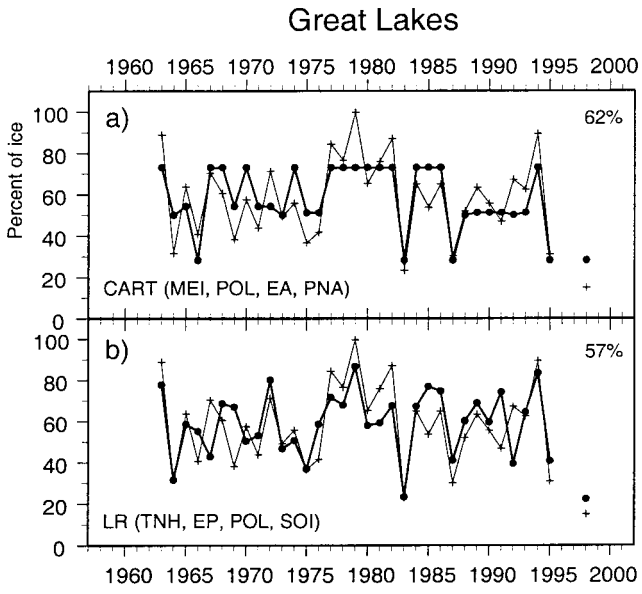

FIG. 9. Same as Figure 4, except for combined ice cover on all the lakes. a) CART model from Figure 3f, b) LR model with TNH, EP, POL, and SOI.

(three splits and four terminal nodes), the R^2 increases to 53% and surpasses that for the LR model. One additional split by CART on NAO was done to separate a group of seven very mild winters in one node. Overall, the tree has approximately even distribution of winters among nodes and describes 58% of the total variance.

Total Ice Cover

The right branch of the tree is the same as for Lakes Superior and Huron, and the first split on the left branch is the same as for Lake Erie (Fig. 3f). One more split on PNA yields 5 terminal nodes with 62% of the total variance. Three of the five terminal nodes have almost the same average values so that the model (Fig. 9a) contains basically three levels: below-, near- , and above- normal. The best LR model (Fig. 9b), ICE_TOT = 0.10TNH – 0.19EP – 0.11POL + 0.05SOI + 0.89, had the same structure as for Lake Huron and describes 57% of the total variance (Fig. 9b).

DISCUSSION

One of the most appealing features of the CART method is its ability to provide an easy and meaningful interpretation of statistical relationships. Using CART a previous analysis (Assel and Rodionov 1998) was expanded by identifying threshold values of individual indices and combination of indices associated with above and below average ice cover and related atmospheric circulation patterns. These results are portrayed in the form of decision trees (Fig. 3). Another (equivalent) form of presentation is a set of IF-THEN rules (Rodionov and Martin 1996, 1999). Each terminal node of the tree-structured model for Lake Superior (Fig. 3a) can be written as a rule. For example, the second node is equivalent to: IF MEI > –1.1 AND MEI ≤ 0.8, THEN ice cover = 83% (± 17%). This rule and similar rules for other lakes suggest a strong association between Great Lakes ice cover and ENSO (MEI) events in the equatorial Pacific. This association, noted by Assel (1998) and Assel and Rodionov (1998), is further explored here.

Another index important for modeling Great Lakes ice cover is TNH. This index was at the root node of the CART model for Lake Michigan (Fig. 3c) and the second node for Lake Ontario (Fig. 3e). Without the MEI index, the TNH index would be the best choice on all the lakes, except Lake Erie. Since the TNH index significantly correlates with ice cover on Lakes Superior, Michigan, Huron, and total ice cover at the 99% level (Table 3) and with Lake Ontario at the 95% level, it is not surprising

TABLE 3. Correlation coefficients (multiplied by 100) between ice cover and atmospheric teleconnection indices.

Index	All Lakes	Superior	Michigan	Huron	Erie	Ontario
NAO	–6	10	–23	7	–7	–15
EA	–34	–32	–15	–22	–35	–3
WP	**–43**	**–51**	–28	–35	–32	–18
EP	–35	–30	–27	–39	–6	**–44**
PNA	–11	–22	7	–3	–21	3
TNH	**47**	**53**	**50**	**46**	23	36
POL	–29	–33	–34	–14	–27	–24
SOI	37	41	21	32	**45**	13
MEI	–37	**–46**	–19	–34	**–46**	–10

Coefficients significant at the 99% level are indicated by bold type. Coefficients significant at the 95% level are indicated by shading. The bootstrap method was used to estimate significance of coefficients. (Efron and Tibshirani 1993)

that this index was chosen first in the LR models for those lakes. Apparently these two indices—MEI and TNH—are the teleconnections associated with interannual variations of ice cover in the Great Lakes. Below their role is considered in detail.

The MEI

The MEI index was used to split the root nodes (Fig. 3) for Lakes Superior, Huron, Erie, and total ice cover. Except for Lake Erie, the splitting point was the same. If MEI is greater than 0.8 (strong El Niño events) nine winters have much below average ice cover: 1964, 1966, 1973, 1983, 1987, 1988, 1992, 1995, and 1998. On Lake Superior, for example, the average ice cover during those winters was only 39%, or 30% below the overall average. For Lake Erie, CART (Fig. 3d) placed the two strongest El Niño events (1983 and 1998) in a separate node.

A composite map of 700-hPa height anomalies for the nine winters with MEI index greater than 0.8 (Fig. 10a) shows a typical response of the Northern Hemisphere circulation to El Niño events during the past three decades. The Aleutian low is much deeper than normal and coupled with a positive anomaly in the subtropical latitudes. This implies strong north-south gradients in geopotential heights and hence vigorous zonal circulation over the North Pacific. A characteristic feature of atmospheric circulation over North America is a positive 700-hPa anomaly centered over the Great Lakes. It coincides with the area of strongest correlation between Great Lakes ice cover and 700-hPa heights over the Northern Hemisphere (Assel and Rodionov 1998). Positive 700-hPa height anomalies in this area indicate that a climatological (quasi-stationary) upper atmospheric trough over eastern North America is less developed, and the Polar jet stream is more zonally oriented. Under this type of circulation surface air temperature anomalies are positive over much of the continent including the Great Lakes basin (Fig. 10b).

As the CART models for Lakes Superior (Fig. 3a), Huron (Fig. 3b), and total ice cover (Fig. 3f) suggest there is still a significant variability in ice cover within this group of nine El Niño winters. Winters in the Great Lakes basin are especially mild, and the amount of ice cover is minimal if El Niño events are accompanied by strong zonal atmospheric circulation over the eastern North Atlantic. The models describe this as MEI > 0.8 and EA > 0.6. If, however, the EA index is less than or equal to 0.6, ice cover is close to normal.

Although ice cover tends to be above normal during non-El Niño events (MEI ≤ 0.8), the overall relationship appears to be non-linear. As the CART model for Lake Superior demonstrates (Fig. 3a), three mild winters (1971, 1974, and 1976) occurred during strong La Niña events (MEI ≤ –1.1). While the 1999 winter was not included in this analysis (because the data on ice cover are not fully processed yet), most likely it will also fall into this category.

A similar asymmetry (nonlinearity) was noted in a number of recent studies of ENSO effect on seasonal precipitation, surface temperature, and teleconnection patterns (Zhang *et al.* 1996, Livezey *et*

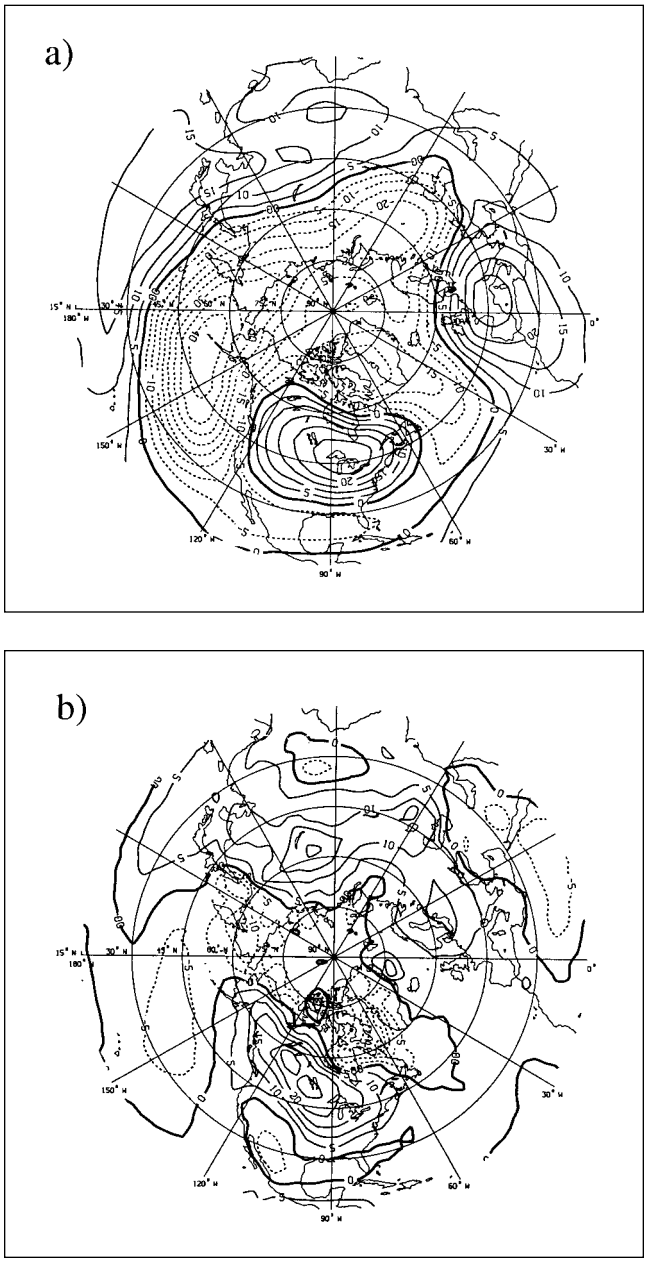

FIG. 10. Composite maps of a) 700-hPa height and b) surface air temperature anomalies for the winters (DJF) when the MEI index is greater than 0.8. The winters are 1964, 1966, 1973, 1983, 1987, 1988, 1992, 1995, and 1998 (9 winters).

al. 1997, Montroy *et al.* 1998). For example, in Mo *et al.* (1998) the response of the WP circulation pattern to the ENSO signal is stronger during El Niño events than during La Niña events. Composite wintertime (DJF) Surface Air Temperature (SAT) anomaly maps over North America for El Niño and La Niña (Hoerling *et al.* 1997), each an average of nine cold or warm events between 1950 and 1996, show a maximum warm temperature anomaly during El Niño is located near Lake Superior. Lake Superior resides near the zero temperature anomaly line of the La Niña composite. The nonlinear component of the SAT anomalies reaches its maximum over the Great Lakes, suggesting separate treatment here for the North American climate response to warm and cold events.

The TNH

The response of ice cover to the TNH circulation pattern appears to be more linear than to ENSO events. It has been noted that there are relatively high correlation coefficients between ice cover and the TNH index (Table 3). Also, the TNH index splits ice cover on Lake Michigan into two roughly equal groups of winters, and the splitting point is close to zero (Fig. 3c). The same splitting point is for this index at the second node of the tree for Lake Ontario (Fig. 3e). Moreover, forcing the TNH index to the root node for Lake Ontario would split exactly the same two groups of years as for Lake Michigan.

The TNH teleconnection pattern consists of two primary anomaly centers: one is centered over the Gulf of Alaska and another one of opposite sign is over the Hudson Bay. This is clearly seen on the composite maps for 19 winters when the TNH index was equal to or below 0.2 (below normal ice) (Fig. 11a), and for 15 winters when the index was greater than 0.2 (above normal ice) (Fig. 11c). Both figures show that the TNH pattern significantly controls the strength and position of the Hudson Bay low and hence southward transport of cold Canadian air into the north-central United States and the Great Lakes region. When the Hudson Bay low is weak and cyclonic activity is enhanced in the Gulf of Alaska, the winds over North America have a significant northward component. As a result, surface air temperatures are above normal over much of the continent (Fig. 11b). Conversely when the Hudson Bay low is strong and cyclonic activity in the Gulf of Alaska is suppressed, frequent outbreaks of cold air from the north keep temperatures below normal (Fig. 11d).

The similarity between the composite maps in Figures 10a and 11a suggests a relationship may exist between the occurrence of a pronounced negative TNH pattern with warm ENSO winters (Barnston *et al.* 1991). The 1997/98-winter season

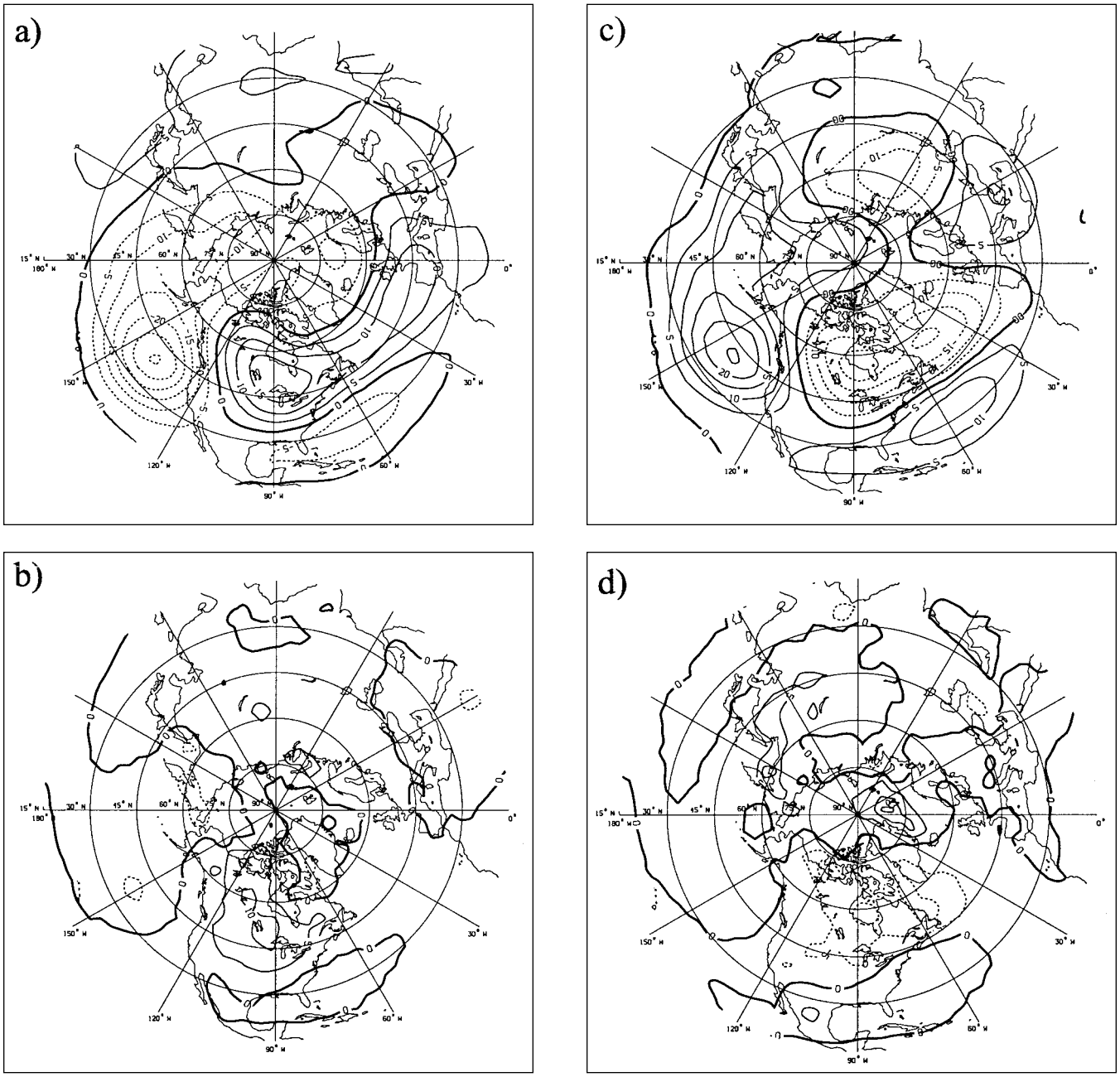

FIG. 11. Composite maps of a, c) 700-hPa height and b, d) surface air temperature anomalies for the winters (DJF) when the TNH index is less or equal than 0.2 (left panel) and greater than 0.2 (right panel). The winters for the left panel maps are 1964–70, 73, 74, 78, 80, 81, 83, 87, 88, 92, 93, 95, 98 (19 winters). The winters for the right panel maps are 1963, 71, 72, 75–77, 79, 82, 84–86, 89–91, 94 (15 winters).

provides a specific example, a strong warm ENSO occurred concurrent with a strong negative TNH pattern. Ice cover on the Great Lakes was at the record low for the period since 1963 (Assel *et al.* 2000).

The CART model for Lake Michigan (Fig. 3c) suggests the hypothesis that the effect of El Niño on ice cover may depend on the phase of the TNH pattern. When TNH is mostly negative and El Niño is strong ice cover is below normal and conversely

with positive TNH and weak El Niño, ice cover is above normal. This relationship can be written as IF-THEN rules:

$$\text{IF TNH} \le 0.2 \text{ AND MEI} > 0.8$$
$$\text{THEN ice cover} = 21\% (\pm 6.4\%), \quad (1)$$

$$\text{IF TNH} > 0.2 \text{ AND MEI} > 0.4$$
$$\text{THEN ice cover} = 91\% (\pm 8.5\%). \quad (2)$$

The first rule describes a group of nine winters with much-below-normal ice cover that went to the right branches of the trees for Lake Superior (Fig. 3a) and Huron (Fig. 3b). The second rule characterizes three cold winters on Lake Michigan (1977, 1979, and 1994) with an average ice cover of 91%. It is important to note that the nine winters in the first group are all winters of *strong* El Niño events (MEI > 0.8), while the winters in the second group are winters of *weak* (0.4 < MEI < 0.8) El Niño events. This opens a possibility for another hypothesis that the effect of El Niño on ice cover may be different for strong and weak El Niño events. Admittedly, the statistical significance of the above rules is low, and the hypotheses require further rigorous testing, preferably on a new independent data set. A detailed analysis of the relationship between ice cover and ENSO events goes beyond the scope of this paper. Here it was important to demonstrate the ability of CART to generate new hypotheses and show the direction where further research is most needed.

SUMMARY AND CONCLUSIONS

Linear regression makes certain assumptions about the data, which can lead to spurious high correlations, or hide real relationships. For example, the assumption that an independent variable affects a whole data set the same way doesn't always hold for variables like ice cover. In the Great Lakes, ice cover usually fluctuates at a certain level, specific for each lake, with rare excursions to extremely low or high values. It is very difficult for LR to model this type of variability. In these tests CART has improved on the accuracy of the LR models by as much as 33% (Lake Erie). Even in those cases where the explained variance of both models was close (Lake Superior), the CART had fewer independent variables than LR. This better performance was achieved largely due to the robustness of CART to the effects of outliers. Extreme values among the independent variables generally affect CART less because it tries to place them in separate nodes.

CART is a nonparametric procedure and does not require specification of a functional form. When the underlying model is unknown and little prior information regarding variable selection is available, CART can help to focus on variables of importance, which makes it a useful exploratory tool. CART's performance can be much enhanced with a researchers knowledge and experience and experiments with various predictors.

A major strength of the CART method lies in its ability to provide insight into physical mechanisms underlying statistical relationships. By presenting these relationships in the form of a decision tree or IF-THEN rules CART facilitates their interpretation. Often this interpretation goes parallel to the process of the model construction and helps to decide whether or not to grow the tree further and which variable should be used for the split.

Two variables – MEI and TNH – were found to be strongly associated with the interannual variations in ice cover. The MEI index was either the best or the second best choice for splitting of the root node of the trees for all the lakes, except Lake Ontario. The analysis of the trees shows that major El Niño events (MEI > 0.8) are accompanied by a significant reduction of ice cover in the Great Lakes. Particularly mild winters are observed when an El Niño event is combined with strong zonal circulation over the East Atlantic. The TNH index was the best choice for Lake Michigan and the second best choice for all other lakes, except Lake Erie. It splits the data into two approximately even groups. Winters in the first group (TNH less than or equal 0.2) have on average much lighter ice cover than winters in the second group (TNH is greater than 0.2).

Since the goal of the CART algorithm is to split the data into homogeneous groups of years, it is natural to combine this method with use of composite maps, a popular statistical technique in climatology. Composite 700-hPa height and SAT anomaly maps were constructed for three groups of winters: 1) MEI > 0.8 (9 winters), 2) TNH ≤ 0.2 (19 winters), and 3) TNH > 0.2 (15 winters). The analysis of these maps has shown that a characteristic feature of atmospheric circulation during winters of below-normal ice cover is a positive 700-hPa height anomaly center over and to the north of the Great Lakes. It indicates a weaker than normal Hudson Bay low, suppressed troughing over eastern North America and more zonal orientation of the jet

stream. On the contrary, during winters with above-normal ice cover, the Hudson Bay low is anomalously deep, which leads to more frequent outbreaks of cold Arctic air.

CART exhibits its greatest advantage over the LR approach when applied to data with a highly nonlinear structure. Many climatic processes are linear only within certain limited interval and become non-linear when the entire range of variation is considered. This appears to hold true for the relationship between ENSO events and Great Lakes ice cover. The CART model for Lake Superior reveals that, although ice cover tends to be below normal during El Niño events, mild winters can also occur during strong La Niña events. Also, the model for Lake Michigan suggests that the effect of strong and weak El Niño events on ice cover may be different. More data are needed to confirm these hypotheses. Nevertheless, even with a relatively small data set CART is capable of uncovering hidden relationships and generating useful working hypotheses.

ACKNOWLEDGMENTS

This is GLERL contribution No. 1209. We wish to thank Klaus Walter for providing the MEI data used in this study. The National Research Council and the National Ice Center supported Sergei Rodionov.

REFERENCES

Angel, J.R., van Dyke, G., and Isard, S.A. 1999. The effect of ENSO on Great Lakes cyclones. In *10th Symp. On Global Change Studies*, pp. 323–326. American Meteorological Society.

Assel, R.A. 1991. Implications of CO_2 global warming on Great Lakes Ice Cover. *Climate Change* 18:377–395.

_____. 1992. Great lakes winter weather-700 hPa PNA teleconnections. *Monthly Weather Review* 120:2156–2163.

_____. 1998. The 1997 ENSO event and implications for North American Laurentian Great Lakes winter severity and ice cover. *Geoph. Res. Letters* 25:1031–1033.

_____, and Rodionov, S. 1998. Atmospheric teleconnections for annual maximum ice cover on the Laurentian Great Lakes. *Int. J. Climatol.*, 18:425–442.

_____, Snider, C.R., and Lawrence, R. 1985. Comparison of 1983 Great Lakes winter weather and ice conditions with previous years. *Mon. Wea. Rev.* 113:291–303.

_____, Janowiak, J.E, Boyce, D., O'connors, C.,Quinn, F.H., Norton D. C. 2000. Laurentian Great Lakes ice and weather conditions for the 1998 El Niño winter. *Bulletin of the American Meteorological Society 81703–717.*

Barnston, A.G., and Livezey, R.E. 1987: Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. *Mon. Wea. Rev.* 115: 1083–1126.

_____, Livezey, R.E., and Halpert, M.S. 1991. Modulation of Southern Oscillation-Northern Hemisphere mid-winter climate relationships by the QBO. *J. Climate* 4:203–217.

Bohanec, M., and Baratko, I. 1994. Trading accuracy for simplicity in decision trees. *Machine Learning* 15: 223–250.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. 1984. *Classification and Regression Trees*. NY: Chapman and Hall.

Burrows, W.R., 1991. Objective guidance for 0–24-hour and 24–48-hour mesoscale forecasts of lake-effect snow using CART. *Wea. And Forecasting* 6, 357–378.

_____, and Assel, R.A. 1992. Use of CART for diagnostic and prediction problems in the atmospheric sciences. In *Proc. 12th Conference on Probability and Statistics in Atmospheric Sciences*, June 22–26, 1992, pp. 161–166. Toronto, Canada, AMS, Boston, USA.

_____, Benjamin, M., Beauchamp, S., Lord, E.R., McCollor, D., and Thomson, B. 1995. CART decision tree statistical analysis and prediction of summer season maximum surface ozone for the Vancouver, Montreal, and Atlantic regions of Canada. *J. Applied Meteorology* 34:1848–1862.

Croley, T.C. 1992. Long-term heat storage in the Great Lakes. *Water Resources Res.* 28:69–81.

_____, and Assel, R.A. 1994 One-Dimensional ice thermodynamics model for the Laurentian Great Lakes. *Water Resources Res*. 30:625–639.

Efron, B., and Tibshirani, R.J. 1993. *An Introduction to the Bootstrap*. New York, New York: Chapman & Hall.

Hartley, S., and Keables, M.J. 1998. Synoptic associations of winter climate and snowfall variability in New England, USA, 1950–1992. *Int. J. Climatol*. 18: 281–298.

Hoerling, M.P., Kumar, A., and Zhong, M. 1997: El Niño, La Nina, and the nonlinearity of their teleconnections. *J. Climate* 10:1769–1786.

International Niagara Working Committee. 1983. *1982–83 Operations of the Lake Erie-Niagara River ice boom.* U.S. Army Corps of Engineers. Buffalo District, Buffalo, New York.

Kushnir, Y., and Wallace, J.M. 1989. Low-frequency variability in the Northern Hemisphere winter: Geographical distribution, structure and time-scales. *J. Atm. Sci*. 46: 3122–3142.

Livezey, R.E., Masutani, M., Leetmaa, A., Rui, H., Ji, M., and Kumar, A. 1997. Teleconnective response of

the Pacific-North American region atmosphere to large central equatorial Pacific SST anomalies. *J. Climate* 10:1787–1820.

Magnuson, J.J., Bowser, C.J., Assel, R.A., Dillin, P.J., Eaton, J.G., Evans, H.E., Fee, E.J., Hall, R.I., Mortsch, L.R., Schindler, D.W., Quinn, F.H., and Webster, K.E. 1997. Potential Effects of Climate Changes on Aquatic Systems: Laurentian Great Lakes and Precambrian Shield Region. *Journal of Hydrological Processes* 11:825–871.

Mo, R., Fyfe, J., and Derome, J. 1998. Phase-locked and asymmetric correlations of the wintertime atmospheric patterns with the ENSO. *Atmosphere-Ocean* 36:213–239.

Montroy, D.L., Richman, M.B., and Lamb, P.J. 1998. Observed nonlinearities of monthly teleconnections between tropical Pacific sea surface temperature anomalies and central and eastern North American precipitation. *J. Climate* 11:1812–1835.

Rodionov, S.N. 1994*: Global and Regional Climate Interaction: The Caspian Sea Experience*. Dordrecht, The Netherlands: Kluwer Academic Pub.

_____, and Assel, R. 1999. Laurentian Great Lakes ice cover and atmospheric teleconnection patterns: A decision-tree analysis. In *Proc. Eight Conference on Climate Variations*, 13–17 September, Denver, CO, pp. 38–43. AMS, Boston, MA.

_____, and Martin, J.H. 1996. A knowledge-based system for the diagnosis and prediction of short-term climatic changes in the North Atlantic. *J. Climate* 9: 1816–1823.

_____, and Martin, J.H. 1999. An expert system-based approach to prediction of interannual variations in the North Atlantic region. *Int. J. Climatol.* 19:951–974.

Rohli, R.V., Vega, A.J., Binkley, M.R., Britton, S.D., Heckman, H.E., Jenkins, J.M., Ono, Y., and Sheeler, D.E. 1999. Surface and 700 hPa atmospheric circulation patterns for the Great Lakes basin and eastern North America and relationship to atmospheric teleconnections. *J. Great Lakes Res*. 25:45–60.

Spear, C., Grieb, T.M., and Shang, N. 1994. Parameter uncertainty and interaction in complex environmental models. *Water Resources Res*. 30:3159–3169.

Wolter, K., and Timlin, M.S. 1998. Measuring the strength of ENSO—how does 1997/98 rank? *Weather* 53:315–324.

Wortley, J.C.A. 1978. *Ice engineering guide for design and construction of small craft harbors*. Advisory Report WSI-SG-78, University of Wisconsin, Sea Grant Communications Office, 1800 University Avenue, Madison, Wisconsin.

Yarnal, B.M., and Leathers, D.J. 1988. Relationships between interdecadal and interannual climatic variations and their effect on Pennsylvania climate. *Ann. Assoc. Amer. Geogr*. 78:624–641.

Zhang, Y., Wallace, J.M., and Iwasaka, N. 1996. Is climate variability over the North Pacific a linear response to ENSO? *J. Climate* 9:1468–1478.

Zorita, E., Hughes, J., and Lettemaier, D.P. 1995. Stochastic characterization of regional circulation patterns for climate model diagnosis and estimation of local precipitation. *J. Climate* 8:1023–1042.

Submitted: 17 April 2000
Accepted: 25 July 2001
Editorial handling: Barry M. Lesht