Synoptic Map-Pattern Classification Using Recursive Partitioning and Principal Component Analysis

ALEX J. CANNON AND PAUL H. WHITFIELD

Science Division, Meteorological Service of Canada—Pacific and Yukon Region, Vancouver, British Columbia, Canada

EDWARD R. LORD

Pacific Weather Centre, Meteorological Service of Canada—Pacific and Yukon Region, Vancouver, British Columbia, Canada

(Manuscript received 5 December 2000, in final form 24 August 2001)

ABSTRACT

A method for classifying synoptic-scale maps into discrete groups is introduced. Tree-based recursive partitioning models are used to develop mappings between synoptic-scale circulation fields and the leading linear and nonlinear principal components (PCs) of weather elements observed at a surface station. Statistically unique but climatically insignificant patterns are avoided by identifying map patterns based on their association with indices related to local weather conditions. The method requires few user-adjustable parameters and includes an algorithm that provides objective guidance for determining the appropriate number of map patterns to retain. The classification method is demonstrated using daily sea level pressure and 500-hPa geopotential height maps from a domain covering British Columbia and the northeastern Pacific Ocean. The linear and nonlinear weather element PCs are derived from daily measurements of surface temperature, dewpoint temperature, cloud opacity, and u and v wind components taken at Vancouver, British Columbia.

Classification performance is tested by applying the method to precipitation and air quality scenarios. Results are compared with those from unsupervised map-pattern classifications based on the k-means clustering algorithm. Results from recursive partitioning models using linear weather element PCs as targets were better than those from the k-means algorithm. Recursive partitioning trees using nonlinear PCs as targets performed slightly worse than those using linear PCs as targets. Interestingly, trees using gridpoint circulation data as inputs outperformed models that used truncated PCs of the circulation data as inputs. Poorer results were found not to result from loss of information due to truncation of the PCs. Instead, the way information is encoded in principal component analysis (PCA) may be responsible for the poor classification performance in the recursive partitioning models using circulation PCs as inputs.

1. Introduction

Synoptic climatology investigates relationships between large-scale atmospheric circulation conditions and the local-scale surface environment (Barry and Perry 1973). Methods in synoptic climatology typically employ one of two fundamentally different approaches (Yarnal 1993). The circulation-to-environment approach structures the circulation data, often by classifying or clustering synoptic-scale maps, prior to seeking links with the local-scale environment. Conversely, environment-to-circulation approaches, such as compositing, structure the circulation data based on criteria defined by the environmental variable. When using the circulation-to-environment approach, expected environmental conditions associated with each map pattern can be calculated. Model performance statistics between observed and predicted values of the environmental variable can then be used to evaluate the strength of the circulation–environment relationship and compare different synoptic climatologies. The environment-to-circulation approach cannot be used in this predictive manner.

While the circulation-to-environment and environment-to-circulation approaches are most common, methods that jointly consider circulation and environmental conditions have been developed. Specification models (Klein 1983) and other forms of empirical downscaling (Hewitson and Crane 1994) seek relationships between synoptic-scale circulation conditions and variables related to the surface environment. In specification, for example, stepwise multiple regression is used to relate gridded circulation data to values of an environmental variable. Despite the fact that the selection of grid points and parameters in the specification equation is not independent of the surface environment, the resulting equation can be used to predict values of the

Corresponding author address: Alex J. Cannon, Science Division, Meteorological Service of Canada—Pacific and Yukon Region, 700–1200 West 73d Ave., Vancouver, BC V6P 6H9, Canada.
E-mail: alex.cannon@ec.gc.ca

environmental variable solely from values of the circulation variables. Model performance statistics between observed and predicted values of the environmental variable can thus be determined. This led Yarnal (1993, p. 128) to conclude that ''specification fits somewhere in the middle between the two approaches to synoptic climatology.''

While powerful, results from specification and other empirical downscaling techniques are quite different than those from techniques based on clustering or classification. Classification-based synoptic climatologies, whether produced manually or via an automated technique, stratify the atmospheric circulation conditions into a series of discrete map patterns. Days belonging to a given group share a common set of properties. The main aspects of the circulation can thus be communicated by a small number of descriptive classes, albeit with some loss of information. Conversely, with specification or downscaling the user obtains an equation that defines some continuous relationship between the circulation variables and a given environmental variable. While the resulting equations may be optimal in a least squares sense and may more accurately describe the continuum of joint circulation–surface states, a discrete description of the atmospheric circulation is not generated. Thus, despite the development of methods that allow the continuous relationships defined by specification methods to be visualized (Hewitson and Crane 1994; Cannon and Lord 2000; Cannon and Whitfield 2001), the results cannot be interpreted or used in the same manner as those from classification systems.

When compared with continuous approaches, automated classification-based methods such as the self-organizing feature map (SOFM) may be better able to describe the continuum of atmospheric circulation states (Cavavos 1999, 2000). Despite being able to provide large, structured classifications with little loss of information, interpretation of the maps can still be difficult without subsequent reclustering of the results (Murtagh 1995). Also, when SOFMs are used to generate smaller classifications [such as those described by Cavavos (1999) and Cavavos (2000)], solutions may not be competitive with standard automated clustering algorithms (Balakrishnan et al. 1994). When compared with specification and downscaling methods, the SOFM is less able to account for linkages between synoptic-scale circulation conditions and local weather conditions. This problem is shared by all automated, unsupervised approaches.

To avoid this problem, a manual classification approach has commonly been adopted. In manually defined synoptic classifications an expert decides how maps are grouped and can ensure that classes have meteorological significance. As stated above, the same is not true of automated, unsupervised map pattern classification methods. While less time consuming to generate and easier to replicate, automated classifications may fail to recognize important links between synoptic-scale circulation conditions and weather conditions at the surface. Such approaches, whether based on correlations, clustering algorithms, rotated principal component analysis (PCA), or newer methods such as the SOFM, all define map patterns based exclusively on synoptic-scale circulation data. Consequently, some of the identified patterns may have little climatic significance; others that occur infrequently but play an important role in controlling aspects of the surface environment may be missed entirely. This is the fundamental weakness of automated, unsupervised map pattern classification methods.

As one solution, Frakes and Yarnal (1997) developed a method combining steps from both manual and automated classifications. Using their method a set of template map patterns is first identified by an expert. Subsequent days are then matched with the templates using an automated correlation-based approach. Frakes and Yarnal (1997) found that the hybrid method was able to reproduce results from a manual synoptic climatology while reducing the time needed to perform the classification. For best results, however, this method still requires a significant time investment to identify representative days suitable for use as templates in the automated classification procedure. An expert must be employed to identify circulation patterns associated with different weather conditions observed at the surface.

Another solution, in this case fully automated, would be to use an unsupervised clustering algorithm with both synoptic-scale circulation data and surface weather element data as inputs. The resulting classification would reflect different synoptic map patterns and local airmass characteristics. The main drawback to this method, however, is that the local weather variables used in the initial classification are required to classify subsequent days. The method could therefore not be used, for example, to stratify GCM fields as only synoptic-scale circulation data would be available as inputs.

As an alternative to traditional manual, automated, and hybrid approaches, a method combining aspects of specification and unsupervised map-pattern classification has been used to produce synoptic climatologies. This method, based on the recursive partitioning model (Therneau 1983; Breiman et al. 1984), is automated, can generate discrete classifications using both circulation and surface weather element data, and can classify new synoptic-scale circulation data without reference to the weather element data. In recursive partitioning models, also known as a classification and regression trees (CART), input variables are related to an output variable using a tree structure; the tree defines a set of rules that act on the input variables such that cases of the output variable are placed into homogenous groups. Because the rules generated by recursive partitioning are simple and easy to interpret, the method has often been used by meteorologists to help produce short-term weather and air pollution forecasts (Burrows 1991, 1997; Burrows et al. 1995; Ryan 1995; Carter and Elsner 1997).

From a climatological standpoint, Faucher et al. (1999) used a combination of tree-based models and fuzzy regression to reconstruct marine winds from large-scale atmospheric predictors. Hughes et al. (1993) and Zorita et al. (1995) first applied recursive partitioning to synoptic classification. In these studies, sea level pressure fields were grouped into discrete map patterns most strongly associated with the presence or absence of rainfall at stations in North America. Schnur and Lettenmaier (1998) used recursive partitioning to generate a synoptic climatology of rainfall states at stations in four regions of Australia. This method was later used by Zorita and von Storch (1999) to classify circulation–rainfall relationships on the Iberian Peninsula.

For application in synoptic climatology, inputs to the recursive partitioning model are measures of synoptic-scale circulation conditions and the model output is the environmental variable of interest. Recursive partitioning results in a set of synoptic classes or map patterns; each individual map pattern is associated with a value of the environmental variable. Like specification, the user obtains a mapping between the circulation variables and local conditions. Unlike specification, however, this mapping can be expressed in terms of a set of discrete circulation types. The strength of recursive partitioning as a synoptic classification technique lies in its use of the local environmental variable as the basis for partitioning the circulation variables. In contrast to automated, unsupervised approaches, joint consideration is given to the synoptic-scale circulation and the surface response. In doing so, circulation patterns relevant to the environmental variable are more likely to be selected; patterns with little significance are more likely to be avoided.

To date, synoptic climatological applications of recursive partitioning have focused on a single local variable of interest. Hughes et al. (1993), Zorita et al. (1995), Schnur and Lettenmaier (1998), and Zorita and von Storch (1999) each related circulation conditions to rainfall. Consequently, the classifications are of little use outside of the original area of application. Studies attempting to relate synoptic-scale circulation conditions to air pollution, damaging winds, water quality, or other environmental indicators would benefit from the results only if the variable of interest were correlated with the target variable originally used to build the model. Unless the target variable is a general indicator of the surface climate at the location of interest, synoptic climatologies derived using recursive partitioning will be of limited use. Ideally, synoptic map-pattern classifications appropriate for a range of variables related to weather at the surface would be developed, rather than having to develop separate classification systems for each variable of interest.

The primary goal of the current study is an extension of recursive partitioning that allows the creation of general synoptic climatologies for a region. To accomplish this, a method that combines recursive partitioning with PCA (Green 1978; Monahan 2000) is proposed. Instead of relating circulation measures to a specific environmental variable, PCA is instead used to generate a one-dimensional (1D) index that reflects contributions from a number of weather-related variables observed at a surface station. This index is then used as a target in the recursive partitioning model thereby generating a classification system that better reflects synoptic-scale controls on surface weather. Results could then be used as an alternative to general manual classifications (Lamb 1972; Muller 1977; Hess and Brezowsky 1977) or automated map-pattern classifications developed for a specific region (Yarnal 1993). Weaknesses of both manual (time consuming to produce and difficult to replicate) and traditional automated (difficult to specify important circulation–surface links) methods are avoided using the proposed technique.

The remainder of the study is split into four sections. First, circulation and surface weather element data used in the synoptic classifications are described in section 2. The recursive partitioning–PCA classification scheme and a benchmark classifier are described in section 3. Performances of the synoptic classifications are evaluated and compared in section 4. Last, study results are discussed in section 5.

2. Data

a. Atmospheric circulation

Gridded sea level pressure (SLP) and 500-hPa geopotential height data from the National Centers for Environmental Prediction–National Center for Atmospheric Research (NCEP–NCAR) 40-Year Reanalysis Project (Kalnay et al. 1996) were used as inputs to the synoptic climatological classifications. Daily averages from 1953 to 1998 were obtained for a region covering Western North America and the North Pacific Ocean (40°–62.5°N, 157.5°–110°W). Data were first smoothed spatially by averaging the 2.5° × 2.5° resolution grids (10 × 20) to 5° × 5° grids (5 × 10). Locations of the grid points are given in Fig. 1.

To reduce the impact of seasonal variability in the magnitude of circulation data on the classifications, moving average filters were applied prior to identification of the map patterns (Hewitson and Crane 1992; Yarnal 1993). For each day in the analysis, gridpoint values were expressed as deviations from a mean value calculated using all data from the 13 days centered on the day of interest. This form of filter preserves spatial patterns in the data but removes variations in average magnitude occurring on timescales longer than 13 days. The 13-day window was selected following Hewitson and Crane (1992) and is based on power spectra and correlograms of the daily SLP and 500-hPa geopotential height time series. The filter therefore prevents classifications from being overwhelmed by seasonal variations in average magnitude while still capturing vari-

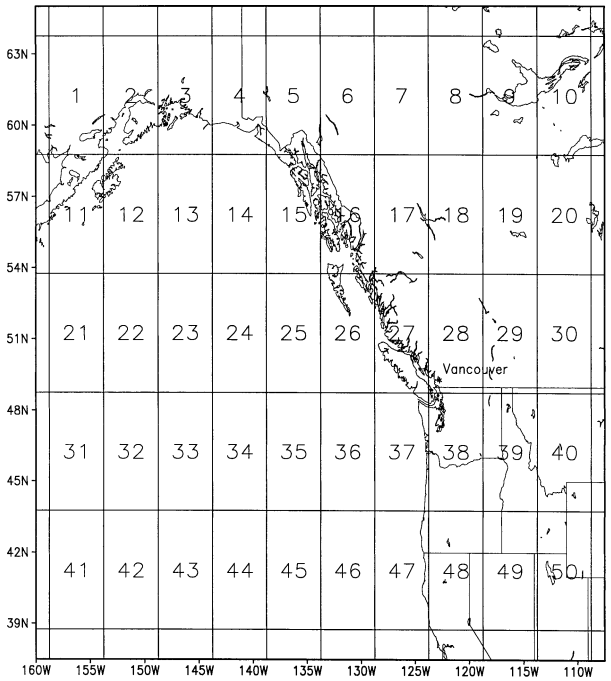

FIG. 1. Map showing the grid points used for the atmospheric circulation data and the location of the surface observation station.

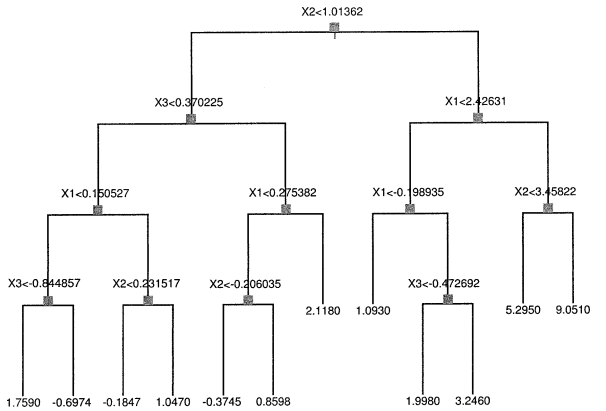

FIG. 2. Example of a recursive partitioning tree with three inputs (X1, X2, and X3). Decision nodes are marked with squares. If the decision rule at a given node is true, cases follow the left branch; if false, cases follow the right branch. The number below each terminal node is the average of the output variable values (Y) for cases assigned to that node.

ability occurring on timescales less than the life span of typical synoptic-scale systems (Hewitson and Crane 1992; Yarnal 1993). Map patterns that reflect only gross seasonal features of the atmospheric circulation are thereby avoided.

b. Surface weather elements

Five variables were selected to describe surface weather conditions in the Vancouver region. Hourly observations of surface temperature, dewpoint temperature, percent cloud opacity, wind speed, and wind direction were obtained from Environment Canada for the monitoring station located at Vancouver International Airport (Fig. 1). These five variables were chosen to reflect standard observations reported by Environment Canada's hourly airport observing stations and therefore allowed simple extension of the procedure to other locations. Variables selected were similar to those used in other studies that defined synoptic types based on local airmass characteristics (Yarnal 1993; McGregor and Bamzelis 1995; Kalkstein et al. 1996; Greene et al. 1999). Local pressures at Vancouver were excluded because gridded SLP data were a part of the atmospheric circulation dataset. This ensured that the two sets of data were independent. For the remaining variables, daily mean values for the period 1953–98 were calculated from the hourly observations. Mean wind speeds and circular mean wind directions were converted into u and v wind components (east–west and north–south, respectively). Similar to the filtering applied to the circulation data, effects of seasonality were removed from the surface weather element data by expressing variables as deviations from centered 13-day moving averages (Yarnal 1993).

3. Method

a. Recursive partitioning

Recursive partitioning is a data-driven statistical model capable of representing nonlinear and interactive relationships between input variables and an output variable. Burrows et al. (1995) provide an excellent description of the model from a meteorological forecasting perspective. Schnur and Lettenmaier (1998) describe recursive partitioning as it applies to synoptic classification. The methodology used in the current study is similar to that presented by Schnur and Lettenmaier (1998), differing in the data type of the output variable. In their description of the method, the output is binary; in the current study the output is continuous. While the variable type determines the criteria used to build the tree, the basic elements of the partitioning procedure are common to the different data types. The following discussion first describes the structure of the recursive partitioning model and then gives a brief overview explaining how these models are constructed. For a complete description of the recursive partitioning algorithm the reader is referred to Therneau (1983), Breiman et al. (1984), and Therneau and Atkinson (1997).

The goal of the recursive partitioning model is to separate the input space in such a way that the output variable cases are placed into groups that are as homogenous as possible. As shown in Fig. 2, the partition is represented using a treelike structure. Inputs to the model are presented at the top of the tree and criteria determining which branch each case proceeds to are made at decision nodes. A question is asked at each

decision node that splits the remaining cases into two groups. Depending on whether the decision criteria is true or false, cases either follow the left branch or the right branch out of the decision node. Cases are assigned to classes based on the terminal nodes they reach in the tree; each terminal node defines a potential class in the synoptic climatology.

Predicted output values for classes are also assigned by the recursive partitioning model. Output values for cases in terminal nodes are averaged; these mean values are used as predictions for the classes. In the tree shown in Fig. 2, for example, a case with inputs X1 = 1, X2 = 2, and X3 = 3 would yield a predicted output Y = 3.246, the mean value of all cases assigned to that terminal node during training. Counting from left to right, this case would be a member of class 10. Residual errors associated with the tree are given by the difference between the observed outputs and the corresponding predicted values. If the observed value were equal to 3, the residual error for this example would be −0.246.

Trees are built using an algorithm that selects and creates decision nodes so that output variable cases are placed into increasingly similar groups. The decision rule at each new node is chosen by iteratively searching through the input variables to find the split that maximizes a measure of node homogeneity. For trees with a continuous target variable, the splitting-criterion SC is given by

$$\mathrm{SC} = \mathrm{SS_T} - (\mathrm{SS_L} + \mathrm{SS_R}), \quad (1)$$

where $\mathrm{SS_T}$ is the sums of squares for the node (equal to the summed squared residual error defined above), and $\mathrm{SS_L}$ and $\mathrm{SS_R}$ are the sums of squares for the left and right branches, respectively. Choosing the highest value of SC leads to the split that maximizes the sums of squares between the new branches.

The algorithm that creates nodes is controlled by two parameters: N_s, the minimum number of cases in a node required to attempt creating a split, and N_t, the minimum number of cases in a terminal node. By default, the recursive partitioning algorithm sets N_s to 20 and N_t to $N_s/3$. In synoptic climatological applications where sample sizes are generally quite large, final tree structure is not very sensitive to these parameters; changes are instead reflected in computation time. New nodes are created until each terminal node contains a minimum number of cases or no further splits can be made because the splitting criterion has converged.

By default, tree size is not limited during the initial fitting process; tree building continues until terminal nodes have reached the minimum size defined by N_t or SC has been maximized and no further splits are possible. Since tree size is not limited, models may overfit the training data used to grow the tree and may not reflect the underlying relationships between inputs and outputs. In the most extreme case, when N_t is set to one and each case is allowed to reach its own terminal node, the residual error of the tree on training data will be zero. The model will have memorized both the structure underlying the data and also noise; performance of the model on data not used in the building process may be poor as a result. As a remedy, overfit trees are pruned by removing unnecessary branches from the model, thereby reducing the number of terminal nodes. The pruning step is very important in synoptic climatological applications as the number of terminal nodes determines the number of map patterns.

The amount of pruning is determined by inspecting out-of-sample estimates of model performance calculated using cross validation (Weiss and Kulikowski 1991). In N-fold cross validation, the available data are first split into N equal size bins. Full trees are then built using data from $N - 1$ groups and residual errors for pruned subtrees are calculated using the data remaining in the leftout bin. This procedure is repeated N times for each subtree, rotating the training and leftout bins at each fold of the cross validation. The N error estimates from cross validation are then collected and their mean and standard error (se) values are computed and stored. Following cross validation, the cross-validated errors for the pruned trees are plotted against the number of terminal nodes. An example is given in Fig. 3; in this case, the error values plotted are proportions of unexplained variance for the pruned trees. As the number of terminal nodes initially increases, cross-validated error typically decreases quite rapidly. This is followed by a relatively flat plateau, and, as overfitting of the training data occurs, a gradual increase in cross-validated error. Trees that are close to minimizing the cross-validated error (i.e., those along the plateau in the plot) are likely to perform well on true out-of-sample data. In practice, performance of these models will be very similar; as a result, the simplest is chosen for the sake of parsimony.

Objectively, the smallest tree that is within one se of the minimum is usually selected as the optimum model (Therneau and Atkinson 1997). The original tree built using the full dataset is then pruned using this criterion, commonly referred to as the ''1-SE rule.'' In Fig. 3, for example, the minimum cross-validated error is 0.67 with an se of 0.03; the 1-SE rule would select the tree with an error equal to 0.67 + 0.03 = 0.70, in this case one with only six terminal nodes. In a synoptic climatological context, however, the 1-SE rule may select a tree with more nodes than can be easily interpreted. Instead, the user can choose to prune the tree to a smaller size, sacrificing model performance for a more compact classification. Further guidance on how to best determine the appropriate tree size for synoptic map-pattern classifications using cross-validation error plots is given in section 4.

b. Linear and nonlinear principal component analysis

PCA is a feature extraction method that attempts to characterize the optimal linear structure of a multivariate dataset by fitting a set of orthogonal axes or principal

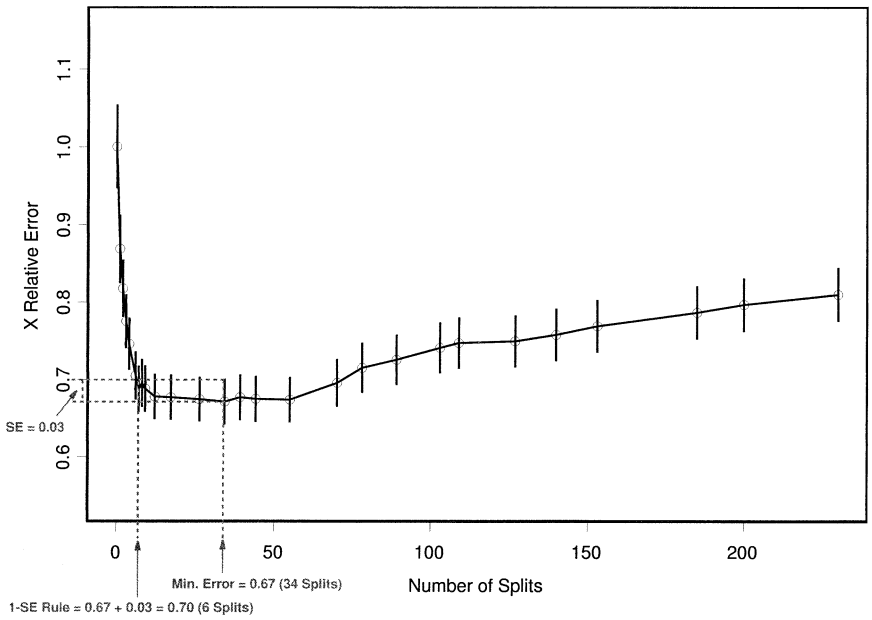

FIG. 3. Example cross-validation error plot obtained during the recursive partitioning procedure. Error bars show ± 1SE. The minimum cross-validated error and the error associated with the 1-SE rule are marked.

components (PCs) through the original data points (Green 1978). Each PC defines a new variable that is a linear combination of the original variables. The first PC accounts for as much variance as possible in the original data; each subsequent PC explains as much of the remaining variance as possible. While the full set of PCs explains all variability in the original dataset, a smaller number of PCs are usually sufficient to explain a large proportion of the total variance. In synoptic climatology, PCA is commonly used as a data reduction tool to characterize large datasets, typically time series of gridded circulation data or observations of surface weather elements, in terms of scores on some smaller number of significant PCs. The PC scores depend both on values of the original observations and on how each PC contributes to the variance in these observations. Because scores on the retained PCs are uncorrelated, they can be used as linearly independent surrogate variables for the original circulation or weather element data. Linear independence and the ability to reduce data dimensionality make PCA well suited for use as a preprocessing step in synoptic classifications based on cluster analysis. In the current study, standard PCA is used to compress the input circulation data and to define a 1D index describing surface weather conditions. Further details are given in the following section.

Standard linear PCA will not achieve optimal results if the underlying structure of the multivariate dataset is nonlinear; alternative methods are required. Described by Monahan (2000), nonlinear PCA (NLPCA) is a generalization of PCA in which a five-layered neural network is used to characterize the nonlinear structure of a dataset. The neural network form of NLPCA reduces to linear PCA if the transfer functions between layers are constrained to be linear. In NLPCA, the first two layers of the neural network define a nonlinear mapping that compresses the input data into a lower-dimensional space defined by the third, bottleneck layer of nodes. Outputs from the bottleneck layer define NLPC scores that are analogous to PC scores in linear PCA. The fourth and fifth layers of the network decode the NLPC scores back to the original dataset dimension and provide a lower-dimensional approximation to the input data. To date, NLPCA has primarily been used as a tool for analyzing modes of variability in climate fields. Monahan et al. (2000) and Monahan et al. (2001) used NLPCA to investigate the leading nonlinear modes of variability in Northern Hemisphere wintertime SLP and 500-hPa geopotential height fields. Monahan (2001) applied NLPCA to sea surface temperature and SLP fields in the tropical Pacific. In each case, NLPCA provided better lower-dimensional approximations to the fields than did linear PCA. As the success of tree-based synoptic classifications in the current study is contingent on the mapping between circulation data and the 1D weather element index, NLPCA may provide better compression of the weather element data and improve classification results. To that end, NLPCA is used as an alternative means of compressing the weather element data and generating the 1D index used as the target for the recursive partitioning model. Further details on the NLPCA procedure are presented in the following section.

c. Recursive partitioning/PCA map-pattern classification

To generate map patterns, the filtered SLP and 500-hPa geopotential height data were used as inputs to a recursive partitioning model. Following the recommendation of McKendry (1994), data from both atmospheric levels were considered simultaneously in the classification procedure. Unlike previous studies that used circulation PCs (Zorita et al. 1995; Zorita and von Storch 1999; Schnur and Lettenmaier 1998), all gridpoint values were used as model inputs in this study. As recursive partitioning algorithms are able to handle large datasets and are not sensitive to correlations between inputs (Burrows et al. 1995), data compression and decorrelation of inputs using PCA are not strictly required prior to classification. In addition, McKendry et al. (1995) suggested that map-pattern classifications based on gridpoint data can be applied to GCM scenarios with greater ease than methods requiring a PCA preprocessing step. Brinkmann (1999) found that information necessary for adequate discrimination between synoptic classes may be contained in higher-order PCs not retained in the PCA. As a result, classification performances of methods that act on the original gridpoint data may be better than those from methods using PCA.

As a test, recursive partitioning models were also built using unrotated S-mode PC scores of the filtered circulation data as model inputs. To ensure equal weighting of grid points from the surface and 500 hPa, PCs were extracted from the correlation matrix of the stacked SLP and 500-hPa geopotential height data. The rule-*N* test (Overland and Preisendorfer 1982) was used to determine the number of PCs to retain. For the 46-yr dataset, 11 PCs were retained representing 93% of the variance in the original set of data. Given the large number of PCs required to represent a substantial fraction of circulation variance, application of NLPCA was not deemed feasible for this particular application.

In a separate step in the classification procedure PCA and NLPCA were used to define 1D indices describing surface weather conditions in Vancouver. Scores of the first PC and the first NLPC from the filtered weather element data were used as outputs in the recursive partitioning models. The linear PC was calculated using a P-mode PCA on the correlation matrix of the filtered weather element data (Yarnal 1993). With one exception, the NLPC was derived following the method suggested by Monahan (2001). Instead of employing a conjugate gradient algorithm, the resilient backpropagation method was used to train the neural networks (Reidmiller 1994). All other steps were followed verbatim. For reference, candidate models were extracted from an ensemble of 20 neural networks, each regularized using stopped training. The maximum number of training iterations was set at 5000. Validation datasets composed 20% of the training data. Four nodes were selected for the encoding and decoding layers; increasing the number of nodes lead to normalized mean square distances between candidates that exceeded the 5% criterion recommended by Monahan (2001). The percent variance explained by the linear and nonlinear PCs is given in section 4.

Following preparation of the circulation data and the weather element indices, recursive partitioning trees were used to generate the synoptic map-pattern classifications. Default values of N_s and N_t (equal to 20 and 7, respectively) were used in the current study. Lowering values of the control parameters did not affect the final pruned trees. Building times, however, increased slightly when N_s and N_t were decreased; small nodes created when control parameters are set to low values are usually irrelevant and are removed during pruning. In the current study, error plots from ten fold cross-validation runs were used as objective guidance for pruning the trees and selecting the appropriate number of synoptic map patterns.

d. Benchmark map-pattern classification procedure

For comparison with the recursive partitioning model, an unsupervised eigenvector-based synoptic map-pattern classification similar to the one described by Yarnal (1993) was also produced. Scores from an S-mode PCA were first calculated from the circulation data and then clustered using a batch *k*-means algorithm (Anderberg 1973). As described in the previous section, PCs were extracted from the correlation matrix of the circulation data and the number of PCs to retain was determined using the rule-*N* test. Cluster centers in the *k*-means algorithm were initialized using the maximum-norm procedure described by Katsavounidis et al. (1994). To test whether or not information contained in higher-order PCs was significant, a separate classification was performed using the full set of filtered gridpoint data. To ensure equal weighting of SLP and 500-hPa geopotential height grid points, data were standardized prior to clustering.

4. Results

a. Weather element PC and NLPC

Prior to performing the synoptic classifications, ability of the PCA and NLPCA to extract information from the weather element data was evaluated. Table 1 shows the fraction of variance r^2 in the filtered weather element data explained by the first PC and the first NLPC. Also listed are fractions of variance in each individual weather element explained by the PC and NLPC, as well as the sign of the correlation coefficients between the filtered weather element variables and the leading linear PC.

The first PC explained 42% of variance in the weather element data. More than 50% of the variance in the temperature variables (positive correlations) and ap-

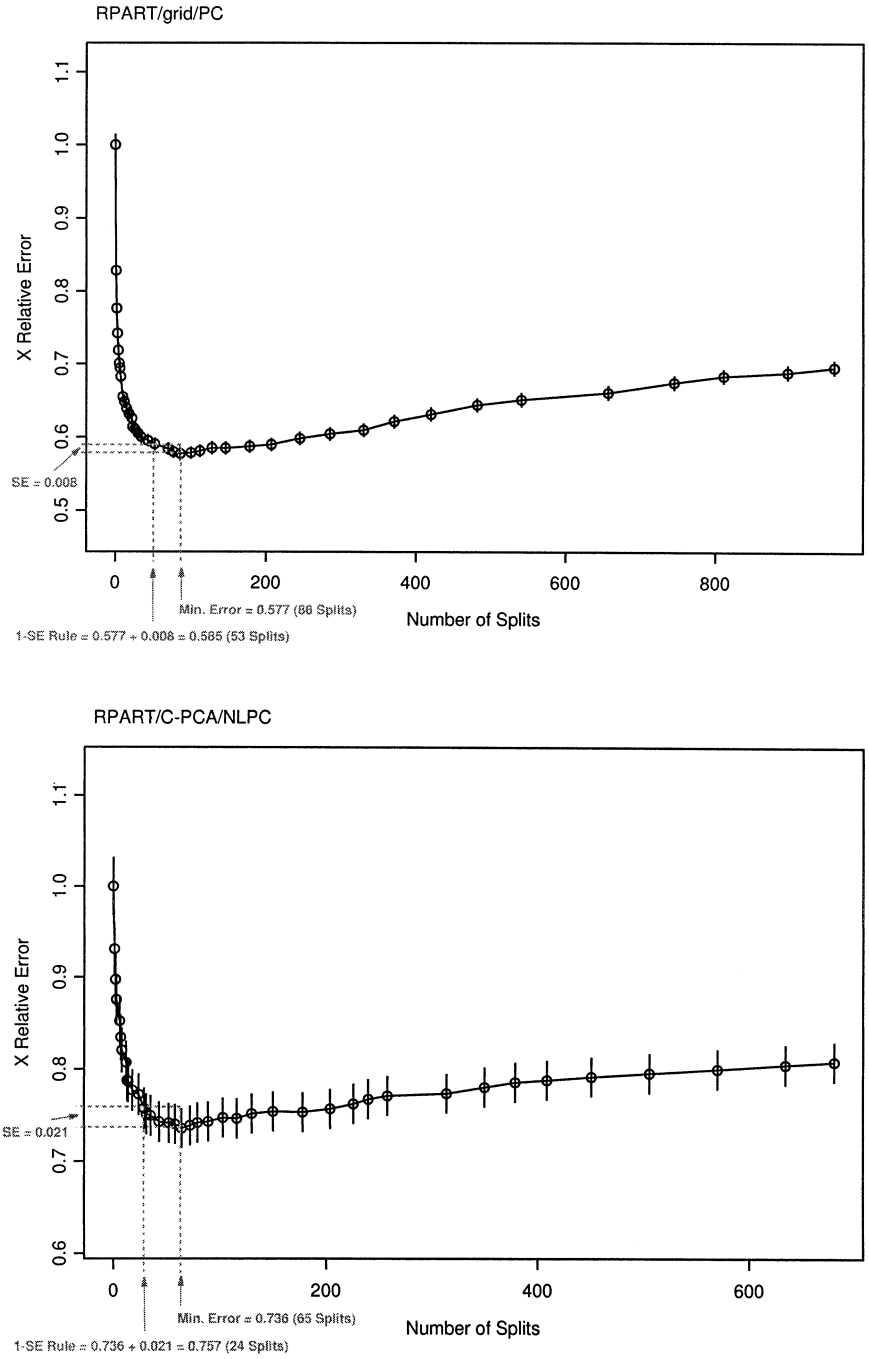

FIG. 4. Cross-validated errors for recursive partitioning trees built in the current study. (top) Results from RPART/grid/PC trees, and (bottom) results from RPART/C-PCA/NLPC trees.

proximately 40% of the variance in the cloud opacity variable (positive correlation) was accounted for by the weather element PC. Slightly less than 25% of the variance in each of the wind components was accounted for by the PC. Scores of the PC were positively correlated with the v wind component but negatively correlated with the u wind component. The first NLPC explained 58% of variance in the weather element data, an additional 16% over the leading linear PC. Increases in explained variance by NLPCA were shown for all variables except the v wind component (2% decrease).

b. Selecting the number of classes

To facilitate comparisons between the recursive partitioning and benchmark models, the number of classes

Table 1. Fractions of variance explained by the first PC and NLPC extracted from the filtered weather element data. The sign of the correlation between each filtered weather element variable and PC1 is shown in brackets.

	PC1 r^2	NLPC1 r^2
Total explained variance	0.42 (+)	0.58
Surface temperature	0.54 (+)	0.88
Dewpoint temperature	0.68 (+)	0.86
Cloud opacity	0.42 (+)	0.48
u wind component	0.23 (−)	0.48
v wind component	0.22 (+)	0.20

was held constant in the current study. The number of classes for all models was selected by inspecting cross-validation error plots for pruned recursive partitioning (RPART) trees (Fig. 4). In this plot, and in the subsequent discussion, models are referenced by combining the classification method (RPART or *k*-means), the type of input data [grid points (grid) or circulation PCA (C-PCA)], and the type of PCA used to compress the weather element data (PC or NLPC). For example, RPART/grid/PC refers to recursive partitioning models with gridpoint data as inputs and the leading linear PC of the weather elements as the output; *k*-means/C-PCA refers to *k*-means classifiers with circulation PCs as inputs.

Plots for RPART/grid/PC and RPART/C-PCA/NLPC models are presented in the top and bottom panels of Fig. 4, respectively; plots for RPART/grid/NLPC and RPART/C-PCA/PC models were qualitatively similar and are not shown. For RPART/grid/PC and RPART/C-PCA/PC trees, the 1-SE rule selected 53 classes and 42 classes with cross-validated errors equal to 0.585 and 0.691, respectively. For models conditioned on the leading NLPC, 28 classes (RPART/grid/NLPC) and 25 classes (RPART/C-PCA/NLPC) were selected with errors equal to 0.656 and 0.757, respectively. As described in section 3, the 1-SE rule can result in a system with too many map patterns to be of practical use. This was true of the models conditioned on the linear PC in the current study. A more generous cutoff of 25 classes, the smallest value recommended by the 1-SE rule for the four RPART models, was chosen instead. This provided classifications of comparable size to well-known historical synoptic climatologies [e.g., 27 weather types by Lamb (1972) and 29 classes by Hess and Brezowsky (1977)]. Also, while subjective, selection of 25 classes lead to classifications that sacrificed little performance relative to those selected using the 1-SE rule. For example, the cross-validated error associated with RPART/grid/PC models using 25 classes was equal to 0.611; this represents a <3% increase in unexplained variance relative to the model recommended by the 1-SE rule.

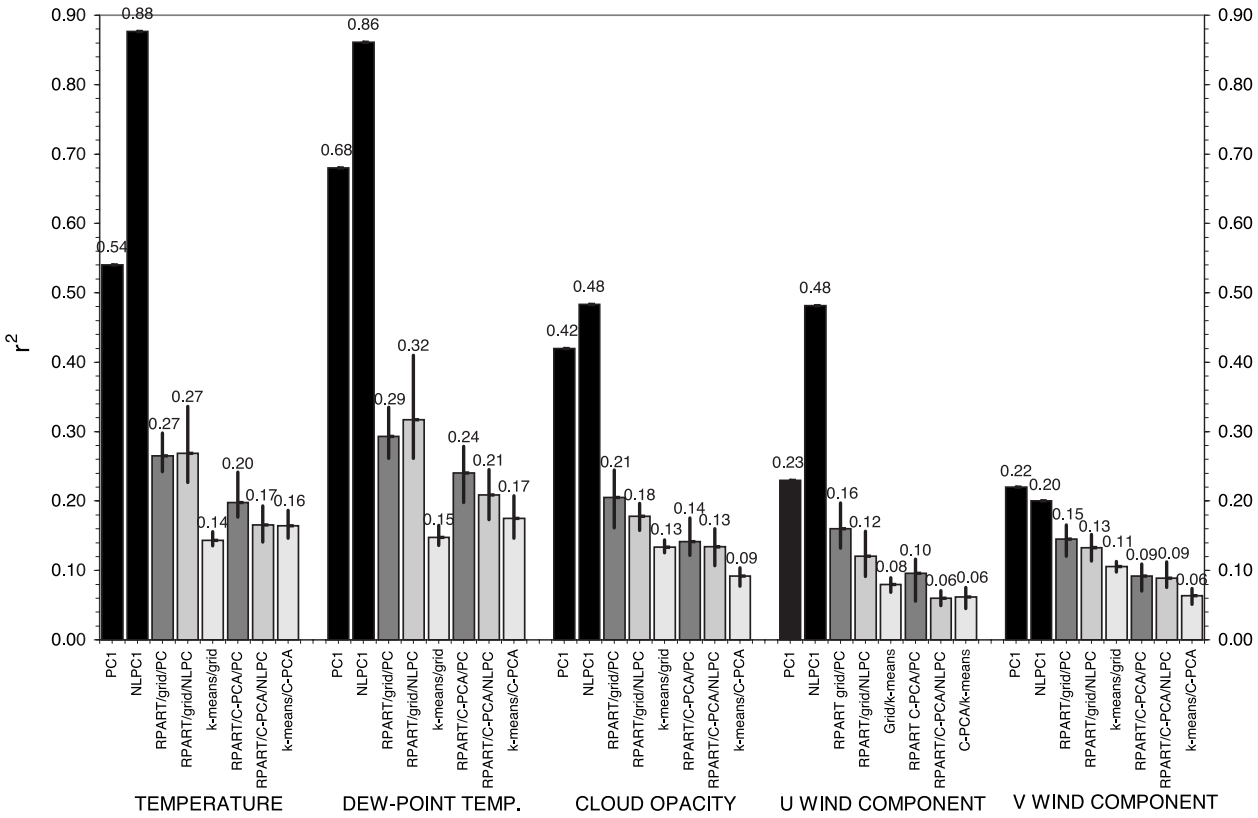

FIG. 5. Mean cross-validated r^2 values for predictions of the filtered weather element data. Error bars indicate the range of r^2 values obtained during cross validation. Values of r^2 for the first weather element PC and NLPC (from Table 1) are shown in black for comparison.

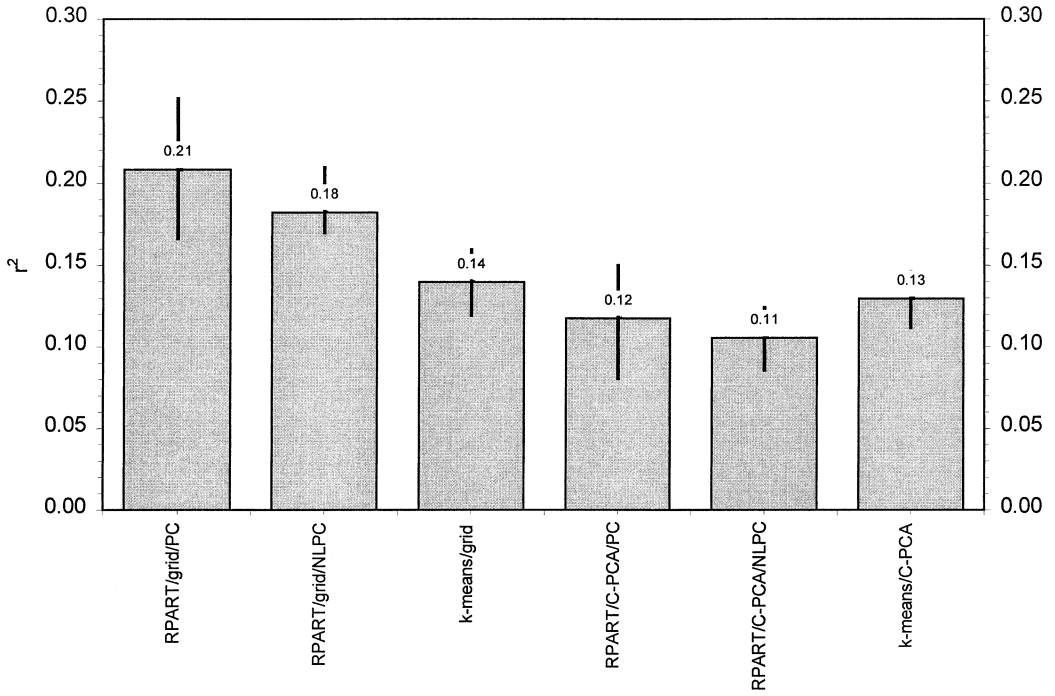

FIG. 6. Mean cross-validated r^2 values for predictions of precipitation at Vancouver International Airport. Lines indicate the range of r^2 values obtained during cross validation.

c. Classification performance evaluation

Synoptic classification performance was evaluated using the method suggested by Yarnal (1993). Each synoptic classifier was evaluated in terms of its ability to predict values of a suite of environmental variables. For each scenario considered, daily values of the observed environmental variable were compared with values predicted by the synoptic classifier. Mean values of the environmental variables were calculated for each of the classes in the worked synoptic climatology; days assigned to a given class then used these mean values for their predicted values.

In the current study, r^2 was used as the primary measure of classification performance. Values of the root-mean-square error and the index of agreement (Willmott 1981) were also calculated, but are not reported due to good agreement with r^2. Relative differences between the models were similar for the three performance measures. To obtain unbiased estimates of model performance, ten-fold cross validation was used to calculate average r^2 values for the period of record. Datasets were randomly split into 10 subsets of equal size. Nine sets were used to generate the synoptic classifications and the remaining set was used to test model performance on data not used in model building. Values of r^2 for the leftout set were recorded and the procedure was repeated, rotating the subsets of data used for training and testing. Reported values of r^2 are means taken over the 10 test subsets. This procedure reduces skill inflation resulting from the evaluation of performance statistics within the dataset used to build the model and values are therefore lower than would be expected if cross validation were not employed.

While cross-validation estimates of classification performance were evaluated in a similar manner as those used to determine pruning of the recursive partitioning models, the two procedures were conducted separately in the analysis. Pruning cross validation was conducted within data reserved for model building and was used exclusively to determine the appropriate number of splits (and thus classes) in the synoptic climatology.

d. Classification performance for the weather element variables

The cross-validation procedure described in the previous section was used to compare the ability of the different synoptic classifications to predict daily values of the five filtered weather elements over the period of record (1953–98). Cross-validated values of r^2 for predictions of each weather element are presented in Fig. 5. For comparison with results shown in Table 1, values of explained variance for the leading PC and NLPC of the weather elements have been added to the plot.

Given that recursive partitioning models used PCs of the weather elements as targets, RPART models were expected to outperform unsupervised classifiers based on the *k*-means algorithm. The gridpoint-based recursive partitioning models (RPART/grid/PC and RPART/grid/

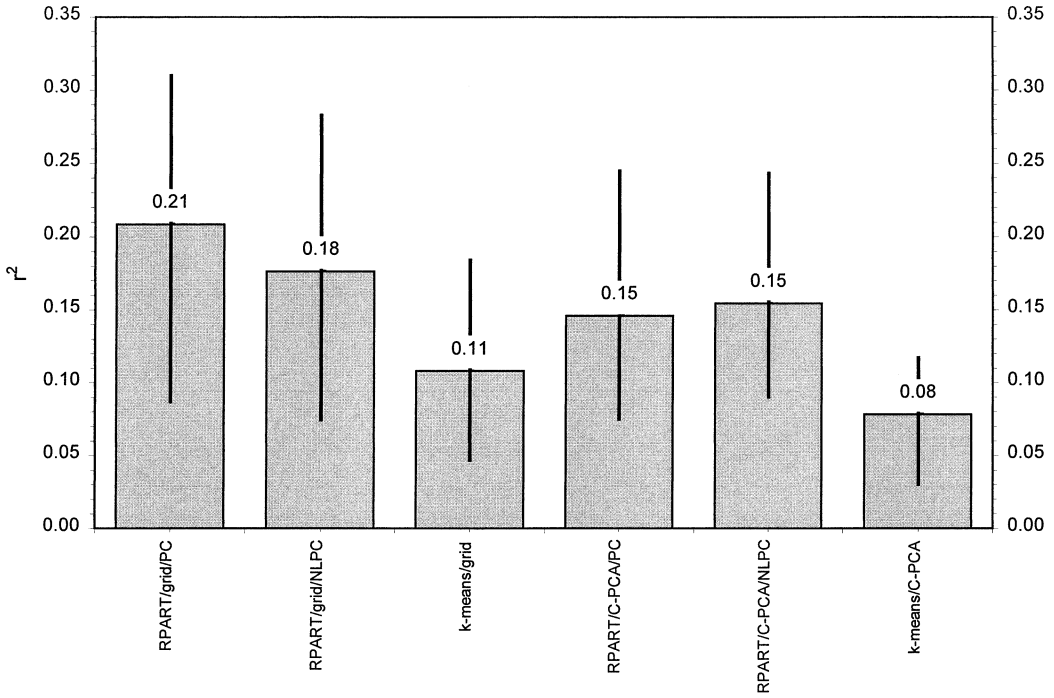

FIG. 7. Mean cross-validated r^2 values for predictions of ozone concentrations at Rocky Point Park. Lines indicate the range of r^2 values obtained during cross validation.

NLPC) did indeed perform better than either of the *k*-means classifications (*k*-means/grid and *k*-means/C-PCA) for all weather elements. Results for the RPART models using circulation PCs as inputs (RPART/grid/PC and RPART/grid/NLPC) were generally better than those for the *k*-means classifiers, but not for all weather elements. The *k*-means/grid model outperformed both of these RPART models on the v wind component and outperformed the RPART/C-PCA/NLPC model on the u wind component.

Models that used gridpoint data as inputs typically performed better than those that used circulation PCs as inputs. This was true for classifications based on recursive partitioning as well as those based on *k*-means clustering. As Brinkmann (1999) suggests, information contained in higher-order PCs may be important when trying to discriminate between map patterns in a cluster-based synoptic climatology. Averaged over the weather elements, the 1% difference in explained variance between *k*-means/grid (12%) and *k*-means/C-PCA (11%) models, however, was within the range of variability exhibited during the cross-validation trials. Conversely, differences in skill between recursive partitioning models using grid points and those using PCs as inputs exceeded the range of cross-validation variability. On average, RPART/grid/PC models explained 21% of the variance in the weather elements versus 15% for the RPART/C-PCA/PC models. Similarly, RPART/grid/NLPC models explained an average of 20% of the variance versus 13% for the RPART/C-PCA/NLPC models.

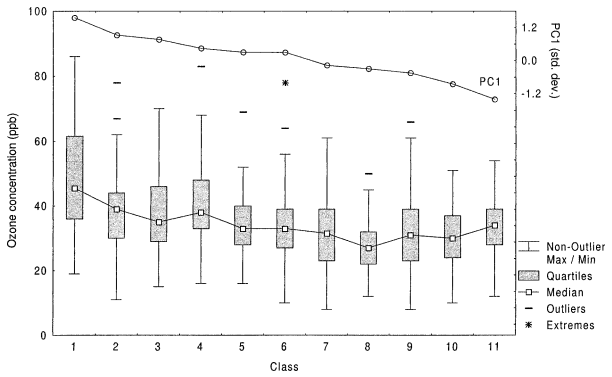

FIG. 8. Boxplots of daily maximum ozone concentrations at Rocky Point Park by RPART/grid/PC classes. Values of the weather element PCs are plotted for reference. Outliers are values greater than 1.5 times the interquartile range. Extremes are values greater than twice the interquartile range.

Information loss due to PC truncation may not have been the sole reason for poor performance of the RPART models that used circulation PCs as inputs. Instead, using gridpoint data as inputs may have offered better discrimination of the local weather element index. Because PCs are linear combinations of the original gridpoint variables, information contained in a PC is distributed across the map area. As a result, each split in the recursive partitioning tree is based on observations at a number of points in space. These points are not necessarily all close to the area where the weather elements are observed. Splits in a tree using grid points

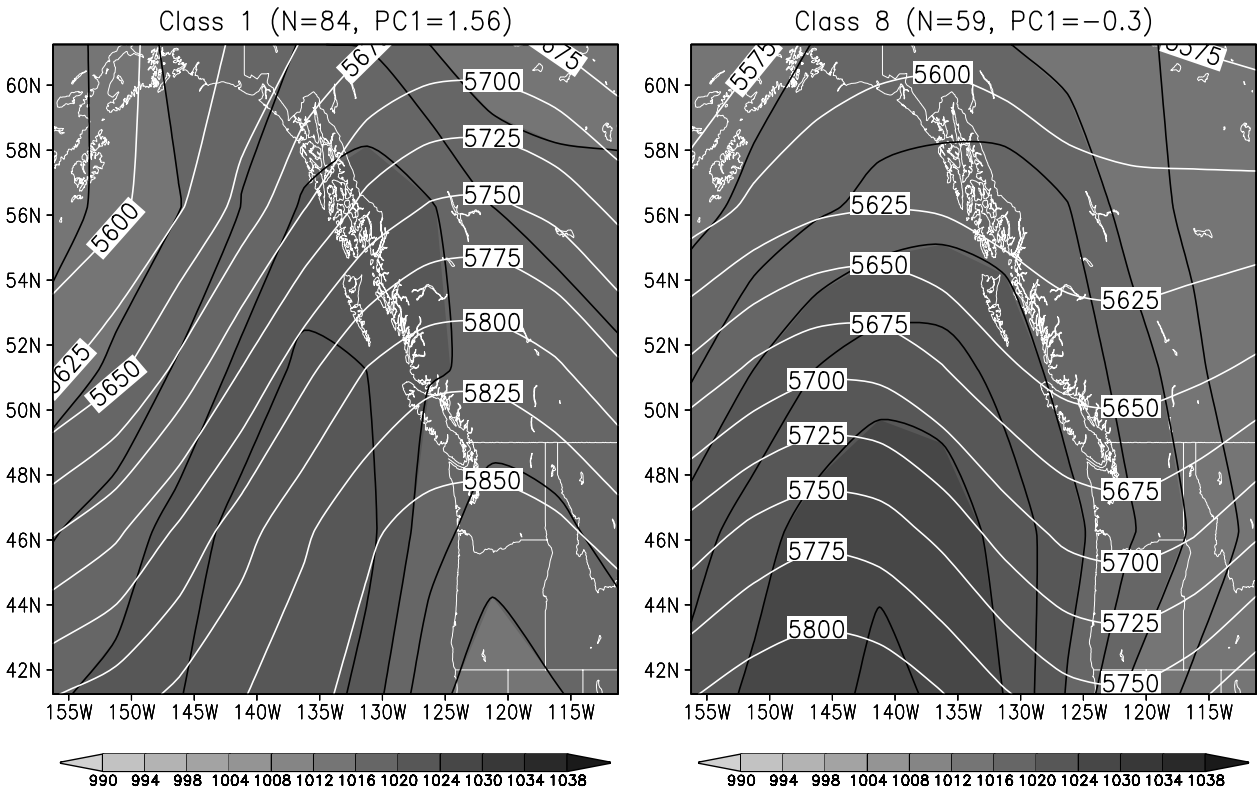

FIG. 9. Example RPART grid/PC map patterns (left) 1 and (right) 8 for the air pollution scenario. SLP field is depicted using dark contours and shading. The 500-hPa geopotential height field is depicted using light contours.

as inputs, however, are based on observations at individual points and can take advantage of circulation information near the local region of interest. To verify, separate RPART/C-PCA/PC models were built using the full set of circulation PCs as inputs. To compare with previous runs, 25 classes were formed using the 1953–98 circulation and weather element data. RPART models using all PCs performed equally well as those using the truncated set of PCs as inputs, but performed worse than those using grid points as inputs. The percentage of explained weather element variance for models using all circulation PCs as inputs was 15%, compared with 22% for those using grid points, and 15% for those using truncated PCs. The PC truncation was not responsible for differences in skill between RPART/C-PCA/PC and RPART/grid/PC models. Instead, gridpoint data were better suited for use as inputs to the recursive partitioning synoptic classification.

While differences were within the range of variability exhibited during cross validation, RPART models conditioned on the linear weather element PC tended to outperform those that used the leading NLPC as a target. Averaged over the five weather elements, a difference of 1% explained variance was noted between the RPART/grid/PC and RPART/grid/NLPC models; the difference was 2% between the RPART/C-PCA/PC and RPART/C-PCA/NLPC models.

e. Classification performance for the precipitation scenario

To assess classification performance on data from outside the set of weather elements, daily precipitation amounts for the 1953–98 time period were extracted from the Vancouver airport station record. The synoptic classifications described above were then used to stratify the daily precipitation observations.

Cross-validated performance statistics from the classifications are shown in Fig. 6. As with the weather elements, RPART models conditioned on the linear PC outperformed those conditioned on the NLPC, although differences were again within the range of variability observed during cross validation. Values of r^2 for RPART models using grid points as inputs and those using circulation PCs as inputs, however, did not overlap during cross validation; improvements for gridpoint-based RPART models were notable. Of the classifiers, the RPART/grid/PC model outperformed all other models, explaining an average of 22% of the variance in precipitation at Vancouver; the RPART/grid/NLPC model explained 18% of the variance. Both gridpoint-based RPART models performed better than the *k*-means models and the RPART models based on circulation PCs. The *k*-means/grid classifier explained 14% of the variance in precipitation, while the three circulation

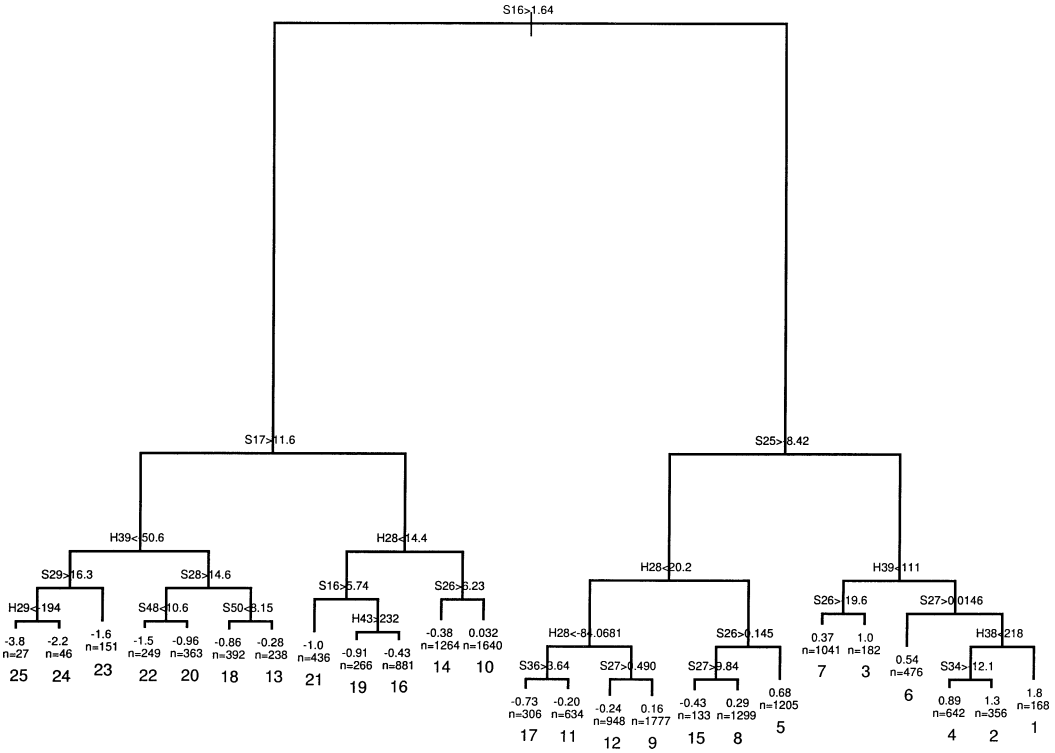

FIG. 10. The 25-class RPART/grid/PC tree obtained during cross validation. If the decision rule at a given node is true, cases follow the left branch; if false, cases follow the right branch.

PCA-based classifiers (RPART/C-PCA/PC, RPART/C-PCA/NLPC, and *k*-means/C-PCA) explained 12%, 11%, and 13% of variance, respectively.

f. Classification performance for the air quality scenario

As another test of the recursive partitioning methodology, separate classifications were applied to summer air quality in the Vancouver area. Previous studies have shown a strong link between surface-level ozone concentrations near Vancouver and synoptic-scale circulation conditions over British Columbia. Taylor (1991) found that above average daily maximum ozone concentrations at a station near Vancouver occurred when a low-level thermal trough and an upper-level ridge of high pressure were present over the region. This relationship was verified by McKendry (1994) using a correlation-based synoptic classification. Pryor et al. (1995) found that 51% of the variance in summer ozone concentrations at the same station could be explained by a linear regression model using circulation conditions at the surface and at the 850- and 500-hPa pressure levels as inputs.

In contrast to the weather element and precipitation scenarios, classifications for air quality were tested using a much smaller subset of available days. This allowed the RPART method to be evaluated in a situation where data were limited. Daily maximum ozone concentrations at Rocky Point Park, a station located just east of Vancouver, were obtained from the Greater Vancouver Regional District's ambient air quality monitoring database. Daily values during late spring and summer (May–September) of 1991–96 were extracted from the database (830 observations total). Using circulation and weather element data from the same 6-yr period, generic synoptic classifications were then constructed and evaluated using the procedures outlined in section 3. Again, SLP and 500-hPa geopotential height data were used as inputs to the classifiers and the first PC and NLPC of the five weather elements were used as targets for the recursive partitioning model. The appropriate number of classes was determined using cross-validation plots and the 1-SE rule. For this scenario, results reported are from models with 11 classes.

Cross-validated performance statistics for the air quality scenario are given in Fig. 7. Patterns of model performance were similar to those shown in the weather element and precipitation scenarios. Due to the reduced sample size, results were more variable than those reported in previous sections; the range in r^2 values for the best performing model and the worst performing model overlapped in this scenario. On average, however, the RPART/grid/PC model again outperformed all other models, explaining on average 21% of the variance in ozone concentrations at Rocky Point Park. The RPART/grid/NLPC model explained 18% of variance, while the

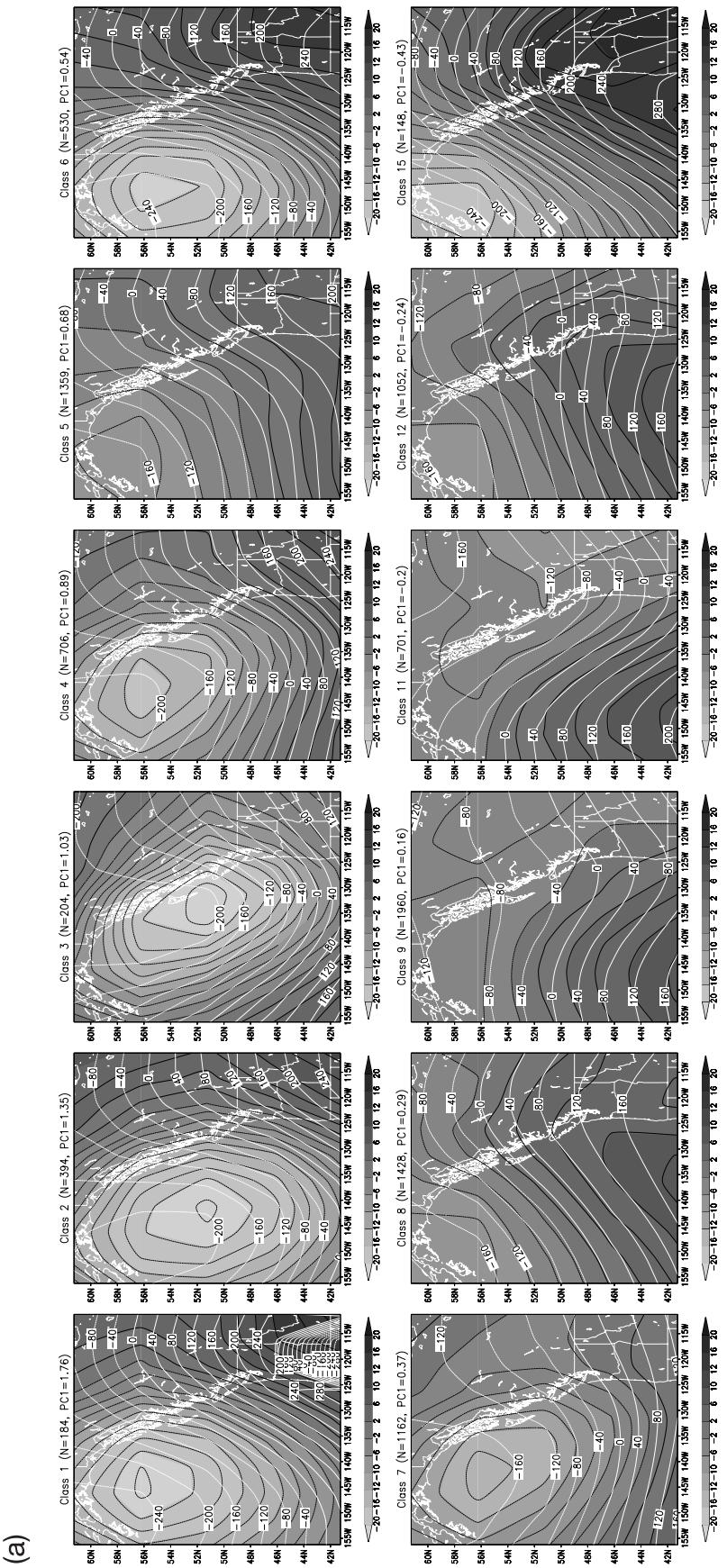

FIG. 11. Composite map patterns defined by the tree shown in Fig. 10. SLP is depicted using dark contours and shading. The 500-hPa geopotential height is depicted using light contours. Composites were formed using all data from 1953 to 1998. (a) Right branch of tree. (b) Left branch of tree.

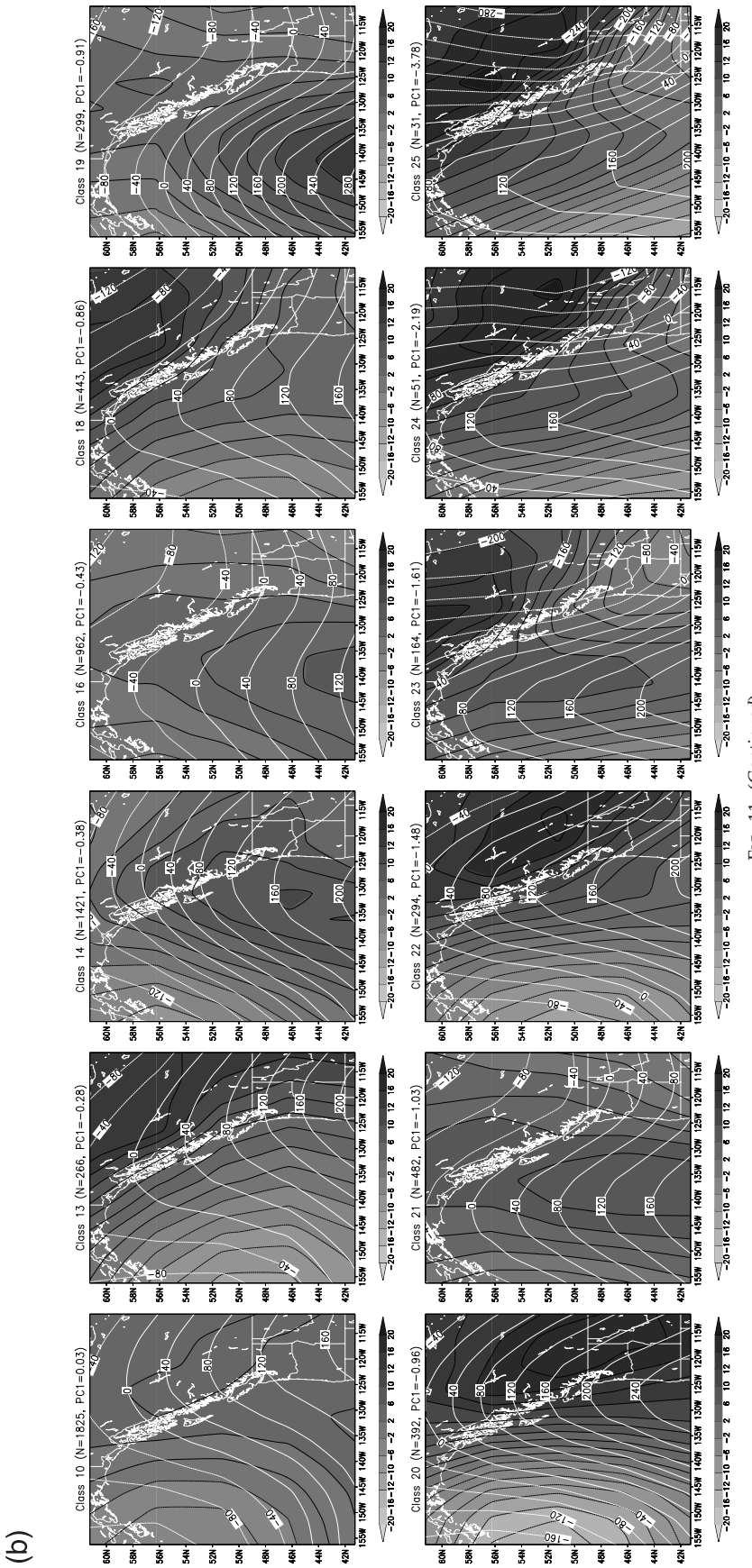

FIG. 11. (*Continued*)

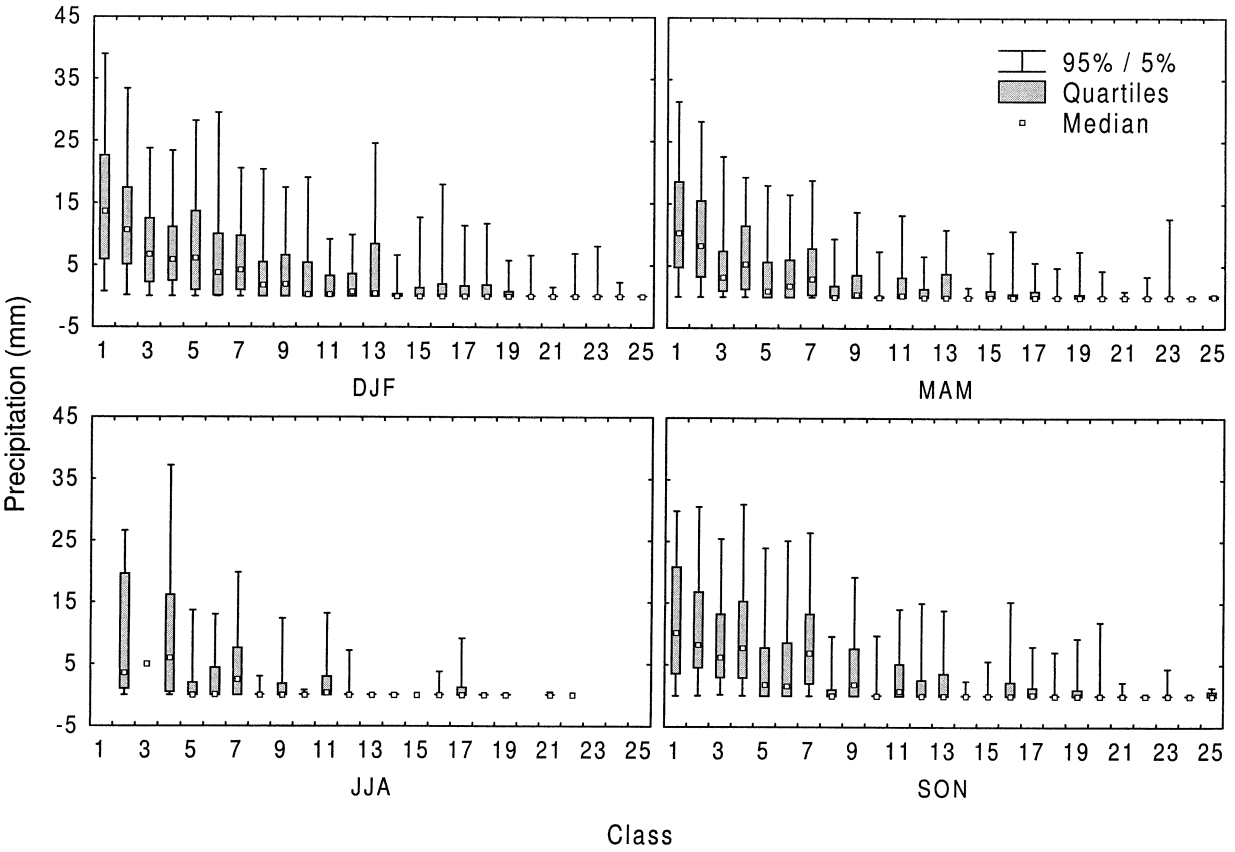

FIG. 12. Boxplots of daily precipitation amounts at Vancouver airport by season and RPART/grid/PC class.

RPART/C-PCA/PC and RPART/C-PCA/NLPC models both explained 15% of the variance. The two *k*-means classifiers (grid/*k*-means and C-PCA/*k*-means) explained 11% and 8% of the variance, respectively.

Distributions of surface-level ozone concentrations in classes defined by one of the RPART/grid/PC models are given in Fig. 8. Of the 80 days with ozone concentrations exceeding 52 ppb, 33 coincided with occurrences of the first map pattern (Fig. 9, left panel). Consistent with findings of Taylor (1991) and McKendry (1994), high-ozone days typically occurred with a strong upper-level ridge between 120° and 130°W and a surface-level thermal trough located along the coast. Lowest ozone concentrations were associated with map pattern 8, shown in the right panel of Fig. 9. In this situation, an upper-level trough was located over the coast, with a surface high pressure center located offshore. Both patterns were linked to below average ozone concentrations by McKendry (1994).

g. Interpreting synoptic climatologies based on recursive partitioning models

In previous sections, synoptic climatologies based on recursive partitioning and standard clustering methods were evaluated in terms of their ability to stratify surface weather elements and environmental variables. Aside from the two map patterns examined for the air pollution example, no attempt was made to interpret output from the map-pattern classifications based on the recursive partitioning model. The following discussion describes how outputs from recursive partitioning can be interpreted and then used to help determine general circulation–surface weather relationships.

An example of one RPART/grid/PC tree obtained during model cross validation is shown in Fig. 10. Variable names above nodes indicate the atmospheric level (S for SLP and H for 500-hPa geopotential height) and grid point from Fig. 1 that each split in the tree was based on. Vertical space between nodes in the plot is proportional to the reduction in residual error associated with the previous split. The first number below a terminal node indicates the predicted value for the weather element PC, the middle value indicates the number of cases assigned to that node, and the bottom number gives the class label for the map pattern. Class labels are sorted in descending order of the predicted PC.

For each terminal node, composite maps have been constructed to show typical filtered anomaly conditions at sea level and at 500 hPa. Map patterns belonging to the right branch of the tree are shown in Fig. 11a; map patterns belonging to the left branch are shown in Fig.

Table 2. Composite values of weather element data for map patterns 1–3 and 23–25 during months (a) Dec–Feb (DJF), (b) Mar–May (MAM), (c) Jun–Aug (JJA), and (d) Sep–Nov (SON). The first number in each column is the composite class mean of the observed weather element variable. The second number represents the composite class mean of the filtered weather element variable (i.e., the mean 13-day anomaly).

Class	No.	Temp (°C)		Dewpoint (°C)		Cloud (%)		u (km/h^{-1})		v (km/h^{-1})	
(a) DJF											
1	108	8.2	3.3	6.2	3.6	92.8	12.9	−14.1	−6.5	7.5	5.2
2	204	7.2	2.4	5.2	2.8	92.2	14.3	−15.2	−6.7	5.4	3.2
3	104	4.1	1.6	1.6	1.7	89.6	10.8	−11.7	−4.1	10.2	7.7
23	94	−2.0	−2.6	−7.5	−4.6	41.0	−22.7	−7.0	−0.8	−1.0	−1.1
24	36	−1.9	−3.1	−9.6	−7.1	28.0	−31.4	−8.5	−2.8	−0.4	−0.4
25	25	−7.3	−6.6	−17.7	−12.2	17.7	−40.9	−11.0	−4.5	1.0	1.1
(b) MAM											
1	26	10.6	2.6	8.0	3.8	87.4	18.1	−11.3	−8.4	6.8	4.3
2	100	9.6	1.6	6.8	2.5	91.2	22.5	−13.5	−8.2	5.3	2.6
3	50	6.3	0.3	2.8	0.5	87.0	18.7	−11.6	−7.7	11.9	9.1
23	18	2.4	−2.1	−3.9	−3.6	48.5	−5.8	−0.8	−0.6	−0.7	−1.2
24	5	1.1	−2.0	−8.4	−6.1	27.3	−28.6	−8.4	−3.4	−6.6	−5.6
25	2	−2.2	−5.0	−9.3	−7.9	68.3	−3.5	−10.6	−11.5	1.5	−1.1
(c) JJA											
1	—	—	—	—	—	—	—	—	—	—	—
2	6	1.62	0.2	12.0	0.9	91.1	37.1	−10.8	−9.1	4.3	2.4
3	1	13.4	−1.3	7.9	−1.9	82.1	10.8	−4.6	−1.4	21.0	13.8
23	—	—	—	—	—	—	—	—	—	—	—
24	—	—	—	—	—	—	—	—	—	—	—
25	—	—	—	—	—	—	—	—	—	—	—
(d) SON											
1	50	10.9	2.9	8.6	3.0	86.6	12.2	−11.0	−5.5	8.7	6.3
2	84	10.9	1.8	8.5	1.8	89.7	15.7	−14.8	−8.0	5.8	4.2
3	49	8.6	1.1	5.7	0.8	89.7	15.4	−14.2	−7.8	8.8	7.0
23	52	2.9	−2.3	−3.0	−4.4	42.0	−20.6	−5.9	−2.1	−3.2	−2.7
24	10	−0.2	−3.4	−7.9	−6.9	16.6	−39.8	−6.5	−2.8	−0.8	−1.0
25	4	−6.5	−5.1	−15.5	−9.2	25.9	−36.3	−11.5	−5.5	1.7	2.6

11b. Combined with information from Fig. 1 and Table 1, general relationships between the synoptic-scale circulation and weather at the surface can be investigated using the tree plot and the composites. As the primary goal of this study was the quantitative assessment of the classification procedure, only brief descriptions of selected map patterns are presented here. Representative of classes from the right and left branches of the tree, map patterns 1–3 and 23–25 are described below. To better show surface weather conditions for these patterns, seasonal class frequencies and mean values of filtered and raw weather element data are given in Table 2.

Based on SLP anomalies north of the Queen Charlotte Islands, the greatest reduction in residual error accompanied the first split in the recursive partitioning model. Composite map patterns in the right branch of the tree (S16 $<$ +1.64 hPa) are typically associated with positive values of the weather element PC, and, as a result, positive anomalies in temperature, dewpoint temperature, cloud opacity, and the v component of wind and negative anomalies in the u component of wind (Table 1). Composite 500-hPa height patterns associated with classes 1–3 are characterized by strong ridging over northwestern North America and a low or trough over the eastern Pacific (Fig. 11a). Moving from class 1–3, the amplitude of the ridge–trough couplet decreases and its location over the region slides southeastward. The upper-level flow over Vancouver is predominantly from the southwest, advecting warm moist air into the region as Pacific frontal systems are steered by the flow toward the north coast of British Columbia or the Alaska panhandle. Vancouver is almost always situated in the warm sector, resulting in cloudy, mild, and moist conditions (Table 2). Surface pressure gradients are strong southeasterly in class 1 patterns and most class 2 patterns, but tend to vary from south through east in class 3 patterns. Winds at Vancouver International Airport tend to be southeasterly and can be quite strong when the cold front lies just offshore.

Patterns in the left branch of the tree (S16 $>$ +1.64 hPa) are generally associated with negative values of the weather element PC, and therefore negative anomalies in temperature, dewpoint temperature, cloud opacity, and the v component of wind and positive anomalies in the u component of wind (Table 1). The 500-hPa height patterns for classes 23–25 are characterized by strong ridging over the eastern Pacific and a low or trough over British Columbia, the Pacific Northwest, or the Great Plains (Fig. 11b). The upper level flow over Vancouver varies from northwest to northeast depending on the location of the 500-hPa low or trough, but it can also be from the south when the low or trough lies over

Vancouver Island. SLP patterns are characterized by high pressure over the interior of British Columbia, low pressure over the Pacific Northwest, and strong pressure gradients over the British Columbia coast. Wintertime weather regimes associated with these patterns are postfrontal. In these situations, cold, dry, arctic air that settled into the British Columbia interior flows out through coastal mountain passes and inlets. At Vancouver International Airport winds are generally from the east to northeast and the weather in the Vancouver region is generally clear, cold, and dry (Table 2).

Once plotted, the map-pattern composites and the tree structure can be used to help interpret relationships between circulation conditions and other environmental scenarios. For the precipitation scenario, for example, boxplots of daily precipitation amounts associated with each class in the RPART/grid/PC model are shown in Fig. 12. Wet and dry classes are very well defined by the classification system, with high precipitation amounts falling into the right branch of the tree and low precipitation amounts into the left branch. As expected, highest precipitation amounts during winter, spring, and fall occurred with map patterns associated with frontal passage (classes 1 and 2). During summer months, class 4, similar to class 1 but with weaker negative anomalies, was associated with the highest median and 95th percentile precipitation amounts. Lowest precipitation totals tended to be associated with the postfrontal conditions (classes 24 and 25).

5. Discussion and conclusions

Used together, recursive partitioning and PCA offer a powerful method for generating synoptic map-pattern classifications. Since classifications are conditioned on a weather element index, resulting classes are more strongly associated with local weather conditions than are automated classifications based only upon synoptic-scale circulation data. For each weather element and scenario reported in the previous section, the best recursive partitioning model outperformed the best unsupervised clustering model.

Unlike unsupervised classifications, the recursive partitioning tree and predicted values of the PC provide an intuitive way of structuring and ordering the resulting map patterns. Variables used to split the data are reported along with the relative importance of each split. In addition, cross-validation error plots generated by the recursive partitioning model can be used to help determine the appropriate number of map patterns, a decision that is often difficult with standard unsupervised cluster analyses. While the 1-SE rule provides a simple, automated criterion for selecting the appropriate number of map patterns for a given dataset, this number can be excessive for large-scale synoptic climatological analyses. Cross-validation error plots can be used to gauge the effect of selecting fewer classes than recommended by the 1-SE rule. Despite this form of guidance, more work is still required to develop automated selection criteria more suited for use with RPART-based synoptic climatologies. An analysis of inflection points in smoothed cross-validation error curves has proved useful and work to develop an alternative criterion based on this methodology is on going.

The success of the classification method presented in the current study depends strongly upon the index used to represent the weather element data. Classification results are strictly valid only for the local region from which the weather elements are drawn. Resulting map patterns are also not ideal for investigating relationships between synoptic conditions and specific environmental variables; for this, targeting the specific variable, as in Hughes et al. (1993), Zorita et al. (1995), Schnur and Lettenmaier (1998), and Zorita and von Storch (1999), would provide better results. Instead, using the weather element index allows general relationships between map patterns and surface weather conditions to be generated for a given region. In this regard, the method provides results that are most similar to manual synoptic climatologies. For example, Maunder's (1968) classification of surface weather maps in the Pacific Northwest was based on links between synoptic conditions and local weather conditions over Vancouver Island. As a result, the classification was less relevant to the interior regions of British Columbia and Washington. The same is true of the classifications in the current study. The fact that the method presented is automated, however, means that it could be easily and quickly applied to a variety of specific locations. Alternatively, data from multiple stations could be combined in the weather element index to create a classification system valid for a larger area.

The usefulness of classifications based on recursive partitioning depends on the strength of the relationship between the variable of interest and the 1D index used to represent the weather element variables. In the current study, weather elements were represented using linear and nonlinear PCs. Despite the improved data compression offered by NLPCA, results for models conditioned on the NLPC were slightly worse than those for models conditioned on the linear PC. Compressing the weather element data down to a single variable, whether using PCA or NLPCA, results in substantial loss of information; for the dataset used in the current study the leading PC explained only 42% of the variance in the original weather elements. While use of this 1D index did result in stronger ties between the synoptic-scale circulation and local weather, it is possible that alternative means of representing the weather elements could lead to even better classifications. For example, clusters from a synoptic airmass typing applied to the weather element data (Yarnal 1993) could be used as targets in a recursive partitioning model. In this case, the circulation data would be split in such a way as to maximize the number of days correctly assigned to each predefined airmass type. This method, however, would

require the number of airmass types to be set prior to partitioning, a step not required when using PCA to define the 1D index. As an alternative, the recursive partitioning model could be modified to allow a multivariate response; this would remove the need for a univariate weather element index altogether and would involve no loss of information due to data compression. Use of multivariate recursive partitioning models for classifying synoptic map patterns is currently under investigation.

Interestingly, recursive partitioning models that used circulation PCs as inputs performed worse than models that used gridpoint data as inputs. Performance gains resulting from the use of gridpoint inputs could not be explained by the loss of information due to truncation of the PCs. Poor performance may instead have been due to the fact that circulation PCs represent contributions from variables observed over a large spatial domain. Splits based on PCs may therefore not provide the best discrimination of weather elements observed at a point location. In the future, both gridpoint and PC representations should be evaluated and the most accurate representation for the particular application selected for use.

Acknowledgments. The authors would like to acknowledge an anonymous reviewer whose thoughtful comments helped improve the original draft.

REFERENCES

Anderberg, M. R., 1973: *Cluster Analysis for Applications.* Academic Press, 359 pp.

Balakrishnan, P. V., M. C. Cooper, V. S. Jacob, and P. A. Lewis, 1994: A study of the classification capabilities of neural networks using unsupervised learning: A comparison with *k*-means clustering. *Psychometrika,* **59,** 509–525.

Barry, R. G., and A. H. Perry, 1973: *Synoptic Climatology: Methods and Applications.* Methuen, 555 pp.

Breiman, L., J. H. Friedman, R. A. Ohlsen, and J. C. Stone, 1984: *Classification and Regression Trees.* Wadsworth and Brooks, 358 pp.

Brinkmann, W. A. R., 1999: Application of non-hierarchically clustered circulation components to surface weather conditions: Lake Superior basin winter temperatures. *Theor. Appl. Climatol.,* **63,** 41–56.

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

——, 1997: CART regression models for predicting UV radiation at the ground in the presence of cloud and other environmental factors. *J. Appl. Meteor.,* **36,** 531–544.

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

Cannon, A. J., and E. R. Lord, 2000: Forecasting summertime surface-level ozone concentrations in the Lower Fraser Valley of British Columbia: An ensemble neural network approach. *J. Air Waste Manage. Assoc.,* **50,** 322–339.

——, and P. H. Whitfield, 2001: Modeling transient pH depressions in coastal watersheds of British Columbia using neural networks. *J. Amer. Water Resour. Assoc.,* **37,** 73–89.

Carter, M. M., and J. B. Elsner, 1997: A statistical method for forecasting rainfall over Puerto Rico. *Wea. Forecasting,* **12,** 515–525.

Cavavos, T., 1999: Large-scale circulation anomalies conducive to extreme precipitation events and derivation of daily rainfall in northeastern Mexico and southeastern Texas. *J. Climate,* **12,** 1506–1523.

——, 2000: Using self-organizing maps to investigate extreme climate events: An application to wintertime precipitation in the Balkans. *J. Climate,* **13,** 1718–1732.

Faucher, M., W. R. Burrows, and L. Pandolfo, 1999: Empirical–statistical reconstruction of surface marine winds along the western coast of Canada. *Climate Res.,* **11,** 173–190.

Frakes, B., and B. Yarnal, 1997: A procedure for blending manual and correlation-based synoptic classifications. *Int. J. Climatol.,* **17,** 1381–1396.

Green, P. E., 1978: *Analyzing Multivariate Data.* Dryden Press, 519 pp.

Greene, J. S., L. S. Kalkstein, H. Ye, and K. Smoyer, 1999: Relationships between synoptic climatology and atmospheric pollution at 4 US cities. *Theor. Appl. Climatol.,* **62,** 163–174.

Hess, P., and H. Brezowsky, 1977: Katalog der Grosswetterlagen Europas (1881–1976). Deutschen Wetterdientes Rep. 115, 249 pp.

Hewitson, B. C., and R. G. Crane, 1992: Regional climate in the GISS GCM: Synoptic-scale circulation. *J. Climate,* **5,** 1002–1011.

——, and ——, 1994: Precipitation controls in southern Mexico. *Neural Nets: Applications in Geography,* B. C. Hewitson and R. G. Crane, Eds., Kluwer Academic, 121–143.

Hughes, J. P., D. P. Lettenmaier, and P. Guttorp, 1993: A stochastic approach for assessing the effect of changes in regional circulation patterns on local precipitation. *Water Resour. Res.,* **29,** 3303–3315.

Kalkstein, L. S., M. C. Nichols, and C. D. Barthel, 1996: A new spatial synoptic classification: Application to air-mass analysis. *Int. J. Climatol.,* **16,** 983–1004.

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.,* **77,** 437–471.

Katsavounidis, I., C. C. J. Kuo, and Z. Zhang, 1994: A new initialization technique for generalized Lloyd iteration. *IEEE Signal Process. Lett.,* **1,** 144–146.

Klein, W. H., 1983: Objective specification of monthly mean surface temperature from mean 700 mb heights in winter. *Mon. Wea. Rev.,* **111,** 277–290.

Lamb, H. H., 1972: British Isles weather types and a register of the daily sequence of circulation patterns, 1861–1971. *Geophysical Memoirs,* No. 116, HMSO, 85 pp.

Maunder, W. J., 1968: Synoptic weather patterns in the Pacific Northwest. *Northwest Sci.,* **42,** 80–88.

McGregor, G. R., and D. Bamzelis, 1995: Synoptic typing and its application to the investigation of weather air pollution relationships, Birmingham, United Kingdom. *Theor. Appl. Climatol.,* **51,** 223–236.

McKendry, I. G., 1994: Synoptic circulation and summertime ground-level ozone concentrations in Vancouver, British Columbia. *J. Appl. Meteor.,* **33,** 627–641.

——, D. G. Steyn, and G. McBean, 1995: Validation of synoptic circulation patterns simulated by the Canadian Climate Centre general circulation model for western North America. *Atmos.–Ocean,* **33,** 809–825.

Monahan, A. H., 2000: Nonlinear principal component analysis by neural networks: Theory and application to the Lorenz system. *J. Climate,* **13,** 821–835.

——, 2001: Nonlinear principal component analysis: Tropical Indo-Pacific sea surface temperature and sea level pressure. *J. Climate,* **14,** 219–233.

——, J. C. Fyfe, and G. M. Flato, 2000: A regime view of Northern Hemisphere atmospheric variability and change under global warming. *Geophys. Res. Lett.,* **27,** 1139–1142.

——, L. Pandolfo, and J. C. Fyfe, 2001: The preferred structure of variability in Northern Hemisphere atmospheric circulation. *Geophys. Res. Lett.,* **28,** 1019–1022.

Muller, R. A., 1977: A synoptic climatology for environmental baseline analysis: New Orleans. *J. Appl. Meteor.,* **16,** 20–33.

Murtagh, F., 1995: Interpreting the Kohonen self-organizing feature map using contiguity-constrained clustering. *Pattern Recognition Lett.,* **16,** 399–408.

Overland, J. E., and R. W. Preisendorfer, 1982: A significance test for principal components applied to a cyclone climatology. *Mon. Wea. Rev.,* **110,** 1–4.

Pryor, S. C., I. G., McKendry, and D. G. Steyn, 1995: Synoptic-scale meteorological variability and surface ozone concentrations in Vancouver, British Columbia. *J. Appl. Meteor.,* **34,** 1824–1833.

Riedmiller, M., 1994: Advanced supervised learning in multilayer perceptions- from backpropagation to adaptive learning techniques. *Comput. Stand. Interfaces,* **16,** 265–278.

Ryan, W. F., 1995: Forecasting severe ozone episodes in the Baltimore metropolitan area. *Atmos. Environ.,* **29,** 2387–2398.

Schnur, R., and D. P. Lettenmaier, 1998: A case study of statistical downscaling in Australia using weather classification by recursive partitioning. *J. Hydrol.,* **212–213,** 362–379.

Taylor, E., 1991: Forecasting ground-level ozone in greater Vancouver and the Lower Fraser Valley of British Columbia. Tech. Rep. PAES-91-2, Scientific Services Division—Pacific Region, Atmospheric Environment Service, 8 pp.

Therneau, T. M., 1983: A short introduction to recursive partitioning. Orion Tech. Rep. 21, Department of Statistics, Stanford University, 11 pp.

——, and E. J. Atkinson, 1997: An introduction to recursive partitioning using the RPART routines. Mayo Clinic Tech. Rep. 61, Section of Biostatistics, Department of Health Sciences Research, Mayo Clinic, 52 pp.

Weiss, S., and C. Kulikowski, 1991: *Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems.* Morgan Kaufmann, 223 pp.

Willmott, C. J., 1981: On the validation of models. *Phys. Geogr.,* **2,** 184–194.

Yarnal, B., 1993: *Synoptic Climatology in Environmental Analysis.* Bellhaven Press, 195 pp.

Zorita, E., and H. von Storch, 1999: The analog method as a simple statistical downscaling technique: Comparison with more complicated methods. *J. Climate,* **12,** 2474–2489.

——, J. P. Hughes, D. P. Lettenmaier, and H. von Storch, 1995: Stochastic characterization of regional precipitation patterns for climate model diagnosis and estimation of local precipitation. *J. Climate,* **8,** 1023–1042.