CART Decision-Tree Statistical Analysis and Prediction of Summer Season Maximum Surface Ozone for the Vancouver, Montreal, and Atlantic Regions of Canada

WILLIAM R. BURROWS

Environment Canada, Atmospheric Environment Service, Meteorological Research Branch, Downsview, Ontario, Canada

MARIO BENJAMIN

Environment Canada, Quebec Region, St-Laurent, Quebec, Canada

STEPHEN BEAUCHAMP

Environment Canada, Atlantic Region, Bedford, Nova Scotia, Canada

EDWARD R. LORD, DOUGLAS MCCOLLOR, AND BRUCE THOMSON

Environment Canada, Pacific Region, Vancouver, British Columbia, Canada

(Manuscript received 16 March 1994, in final form 2 September 1994)

ABSTRACT

Prediction of daily maximum surface ozone (O_3) concentration was begun by Environment Canada in the spring of 1993 for the Vancouver, Montreal, and Atlantic regions in order to advise the public of expected air quality. Forecasts have been issued for southern Ontario for many years by the province of Ontario, but this is a new undertaking in other parts of the country, where air quality has become a concern in recent years. There is a need for guidance to prepare the forecasts, particularly for prediction of surface O_3 concentration levels near or exceeding the Canadian 1-h maximum acceptable concentration of 82 ppb. Such occurrences are episodic and relatively rare in southern Canada. Probability of occurrence is in the range 0.00–0.08 at the sites in the regions studied here, thus, reliable prediction is difficult without guidance. Mesoscale numerical meteorological–photochemical models are not currently available for routine use in operations, but the capability exists for development and use of sophisticated multivariate statistical techniques for prediction of daily maximum O_3 concentration. Most statistical ozone forecast procedures to date in Canada have been based on multiple linear regression with a limited number of predictors mainly drawn from surface meteorology and subjective classification of the synoptic meteorological flow pattern. Since the relationship between surface O_3 and meteorology is nonlinear, tree-based statistical models with several predictors are appropriate for developing objective forecast guidance.

Surface and upper-air meteorological predictors and other predictors were matched with several years of observed daily maximum O_3 concentrations for the months of May–September at air-monitoring sites in the three regions. A recent nonparametric data-driven tree-based analysis method known as CART (classification and regression trees) was used to analyze the data at each site. The decision trees built by CART were found to fit the data reasonably well, and the rules for node splitting were found to be physically realistic. Some of the important aspects of the analyses are noted. One interesting result was that moisture content of the air plays a limiting role on the maximum surface O_3 concentration that can be achieved when other factors point to occurrence of high values.

The decision trees can be used to predict maximum surface O_3 concentrations if the predictor variables are forecast, thus providing an inexpensive site-specific model for forecasts and climate impact analysis. An estimation of performance with independent data was conducted for the Vancouver–lower Fraser River valley and Montreal regions for each of the five years 1988–92. Verification of the ensemble of forecasts in the two regions shows the technique would have reasonably good skill in forecasting surface O_3 concentrations near or exceeding acceptable 1-h limits. A computer version of the technique has been provided for use in the regional forecast offices.

1. Introduction

In elevated concentrations, surface ozone (O_3) has potentially deleterious effects on human and animal health, vegetation, and certain materials. It is a secondary photochemical pollutant produced in reactions involving a variety of natural and anthropogenic precursors, which include oxides of nitrogen (NO_x) and volatile organic compounds emitted by biogenic, industrial, and vehicular sources. High concentrations of surface O_3 are known to be strongly correlated with meteorological conditions that affect photochemical

Corresponding author address: Dr. William R. Burrows, Meteorological Research Branch, Atmospheric Environment Service, Environment Canada, 4905 Dufferin St., Downsview, ON M3H 5T4, Canada.

chemical reaction rates and long-range transport of O_3 and its precursors (Chung 1977; Heidorn and Yap 1986; McKendry 1993; and many others). In the spring of 1993, Environment Canada began daily preparation of air quality forecasts for the Vancouver, Montreal, and Atlantic regions, where air quality has become a concern in recent years. Occurrences of high surface O_3 concentrations have long been recognized in the southern Ontario region of Canada, where the province of Ontario has been issuing air quality forecasts for several years. However, in other parts of Canada where air quality has become a concern in recent years and provincial facilities are not in place, Environment Canada, a federal government agency, has undertaken the responsibility of issuing air quality forecasts. An air quality advisory for affected areas is issued jointly with provincial authorities when the maximum 1-h average concentration of surface O_3 is forecast to equal or exceed 82 ppb. There is a need for development of forecast guidance for use in the operational decision-making process. High-resolution numerical meteorological–photochemical models are not yet available for routine use in operations, but the capability exists for development and use of statistical techniques to forecast surface O_3. Indeed, there is evidence that statistical models will be increasingly called upon to serve alongside numerical models to compensate for inconsistencies in forecasts of many surface atmospheric elements (Ramage 1993). In the Canadian regions affected by surface O_3, there has been a network of air quality monitoring stations operated by federal and provincial authorities for several years, giving daily observations of levels of O_3 and, at some sites, other gases and particulates. An extensive network of upper-air and ground-based meteorological observations covering much of the country for many years are available. Thus, the necessary historical data needed to develop statistical forecast models for surface O_3 prediction is available.

Most previous work in Canada relating surface O_3 to meteorological and other predictors have used multiple linear regression models with a limited number of predictors, mainly drawn from surface meteorology. Robeson and Steyn (1990) tested two types of time-series models and a two-variable (maximum temperature and persistence) linear regression model to predict maximum daily O_3 concentration for the Vancouver area. Although useful in some situations, they found these simple models have limited capability and have difficulty forecasting the relatively rare occurrences of high O_3 concentrations. Multiple linear regression equations relating persistence, cloud, and surface maximum temperature and wind speed were developed for use as forecast tools for the Toronto region of southern Ontario and to estimate the influence of long-range transport to surface O_3 (MEP 1990). In addition to these and other predictors, the province of Ontario's procedure for preparation of an air quality forecast includes an assessment of upstream air pollution conditions in the bordering states of the United States and the influence of upper-air meteorology and long-range transport by subjective classification of the synoptic meteorological flow pattern or "map type" (Heidorn and Yap 1986). Occurrences of high surface O_3 concentrations are relatively rare and episodic in southern Canada since certain optimal meteorological conditions must be met (Yap et al. 1988) and the relationship between surface O_3 and meteorological predictors is nonlinear. Stockenius (1991) noted that decision-tree regression models are appropriate and will have potential advantages over linear regression models, and this is our experience as well.

One of the authors of this paper has developed multivariate data-driven rule-based statistical models for operational forecasts of episodic meteorological elements, which can be useful for forecasts of common and rare events (Burrows 1990, 1991; Burrows and Assel 1992). The models have proven popular with users because of their simplicity and accuracy relative to other methods, their similarity to the thought process a forecaster uses to make decisions, and the ease with which they can be used on a computer. Decision rules are derived from statistical relationships found between the response variable and predictors that minimize the variance in a training database. Much of the pioneering development work was done by Breiman et al. (1984) in their nonparametric data-driven rule-generating analysis procedure called CART (classification and regression trees). CART model structure is similar to the rule architecture in a neural network and can perform equally as well (Atlas et al. 1990). Breiman et al. (1984) found error levels of CART solutions to be nearly always as low or lower than solutions by parametric procedures such as linear regression, logistic regression, and discriminant analysis, and errors are significantly lower for problems involving a complex response variable and many predictors. Both single predictors or linear combinations of predictors can be used to find node-splitting rules, thus CART can combine several important physical factors into a single rule that would be very difficult for a person to formulate unaided. The rules are found to make physical sense provided the modeler has chosen predictors that are known to be well related to the response variable. The ease with which the tree can be interpreted and applied makes CART an important alternative to other procedures, even when its performance is comparable to theirs. Excessive computation is not required to obtain results as finding the best or most suitable trees is straightforward and takes only a few seconds to a few minutes on a workstation. Independent data can be dropped down the trees in negligible time. An early version of CART was used in analysis of surface O_3 and meteorological conditions in southern California (Zeldin and Cassmassi 1978). Recently, Hudischewskyj and Stockenius (1991) and Stockenius (1991)

have used CART to study relationships between ozone episodes and various meteorological predictors and persistence in the southern California and Philadelphia areas of the United States.

To forecast surface O_3, CART decision trees generated from a learning database for each site are used in real-time conditions. The forecaster supplies values for the predictors, some of which are known at forecast issue time and some of which are forecasts valid at some time in the future. The trees are invoked with these predictor values, and a forecast is made based on the terminal node that is reached. For multiple sites the procedure is simplified when several predictors are the same for all sites and a computer program does the calculations. Further simplification of the work is possible by automatically updating many of the predictor values with output from the Canadian Meteorological Centre's operational numerical weather prediction (NWP) model. Computer codes to run the CART models have been provided for use in the three Canadian regions.

Analysis and testing methodology and some results are shown in sections 2 and 3, respectively. The decision trees can be used to forecast the full range of surface O_3 concentrations, but emphasis here is on the analysis and forecasting of O_3 exceedances of at least 82 ppb (at least 80 ppb for Montreal). For space considerations, analysis results are discussed in more detail for the Vancouver region than elsewhere. Summary and concluding remarks appear in section 4. A few comments on the CART regression procedure and some features that are useful for data analysis are given in the appendix.

2. Analysis

Daily maximum O_3 observations and predictors were matched for May–September, when the majority of high ozone concentrations occur in Canada. Separate CART regression analyses were done for each site. Upper-air meteorological data come from the nearest radiosonde station; surface meteorological data come from local stations. There are a few differences between the predictors used in the three regions to accommodate local meteorology and forecaster preferences. The observed maximum O_3 data provided to us for the Montreal region was rounded to the nearest 10 ppb before 1992 for all sites except Roxboro and Dorval, but this is not expected to have a significant effect on results.

a. Vancouver–lower Fraser River valley region

Daily observations of maximum surface O_3 concentration for 1985–92 were analyzed at five sites in greater Vancouver and the lower Fraser River valley. Shown in Fig. 1, this region is a broad, relatively flat valley bounded by mountains on the north, east, and south, and by the Strait of Georgia on the west. There are two

FIG. 1. Vancouver–lower Fraser River valley region. Air monitoring sites are marked with "+"; surface climate data sites with "×." Urban areas are shaded lighter than water areas.

urban sites near the sea: Robson Square, in the heart of downtown Vancouver, and Richmond South in the city of Vancouver. Three sites are some distance inland from the sea. Rocky Point Park is at the end of a long narrow ocean inlet bordered by steep mountains and lies just east of a complex of light industry and oil refineries. Surrey East is in a suburban region just east of the city of Vancouver. Chilliwack is in a small urban setting at the eastern end of the Fraser River valley about 90 km east of Vancouver. None are adjacent to major highways.

Surface O_3 concentrations in the Vancouver–lower Fraser River valley region are primarily controlled by local production and meteorological factors since mountain ranges block off significant cross-boundary transports. Ozone concentrations greater than or equal to 82 ppb occur during summer periods when the atmosphere is stable and the weather is sunny and hot. These conditions are set up when a strong upper-air ridge is present along the Pacific coast of the United States and Canada, allowing hot, dry, stable air to move northward along the backside of the ridge from the southwestern United States (Lord 1993). The hot air lowers air pressure along the coast, forming a surface thermal low pressure trough. The easterly pressure gradient in the trough shuts off the normal sea-breeze circulation that develops on warm summer days and, combined with the trapping effect of the mountains and the stable air, prevents the air from being flushed of pollutants. When these conditions persist, the concentration of surface O_3 can exceed acceptable levels until the upper ridge and thermal trough move inland, allowing fresh air to invade the valley.

Figure 2 shows statistics for daily maximum surface O_3 concentrations, with sites grouped by each region. As seen in Fig. 2a, the number of available observations for the Vancouver–lower Fraser River valley region was 955 at Richmond South and about 1150 at the other sites. Figure 2b shows that average O_3 concentrations for the Vancouver region are lowest at the two urban sites, especially at Robson Square, and increase

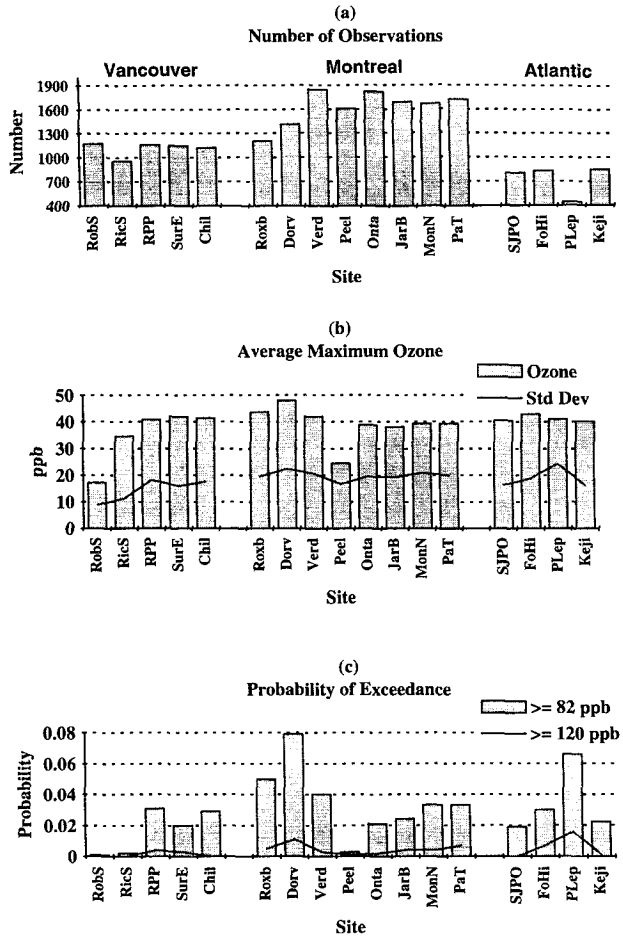

FIG. 2. Summary of O_3 observation data for air monitoring sites, grouped from left to right by region. Locations of sites are shown in Fig. 1 (Vancouver), Fig. 5 (Montreal), and Fig. 7 (Atlantic). All sites are May–September, years are 1985–92 for Vancouver, 1980–92 for Montreal (except Dorval is 1981–92, Roxboro is 1982–92), and 1985–90 for Atlantic (Point Lepreau is 1986–90). Graphs shown: (a) total number of available observations, (b) average and standard deviation of observed daily maximum O_3, (c) probabilities of maximum O_3 of at least 82 ppb (80 ppb for Montreal) and 120 ppb. Name and abbreviation of sites in groups: Vancouver: Robson Square (RobS), Richmond South (RicS), Rocky Point Park (RPP), Surrey East (SurE), Chilliwack (Chill); Montreal: Roxboro (Roxb), Dorval (Dorv), Verdun (Verd), Peel (Peel), Ontario (Onta), Jardin Botanique (JarB), Pointe aux Trembles (PaT); Atlantic: Saint John Post Office (SJPO), Forest Hills (FoHi), Point Lepreau (PLep), Kejimkujik National Park (Keji).

inland where sea-breeze incursions of marine air are less frequent. The very low concentration at the downtown Robson Square site is due to O_3 scavenging by vehicular emissions of oxides of nitrogen, in particular NO, which is common for downtown urban sites (e.g., see McKendry 1993). As portrayed in Fig. 2c, maximum daily O_3 concentrations above 82 ppb are rare events, with probabilities of occurrence only 0.02–0.03 at the inland sites and nearly zero at the two urban sites. Probabilities of occurrence of maximum daily O_3 concentrations of at least 120 ppb are very low, but not zero, at the inland sites and are zero at the two urban sites. The population of Vancouver and the nearby area has been growing rapidly in recent years. It has not been shown if changing emissions patterns have affected O_3 concentrations there, although this may be the case at the Rocky Point Park site (Pryor et al. 1995); however, the effect would not be large in the 8-yr period if this were true.

The predictors used for the Vancouver region appear in Table 1. Most pertain to the surface and upper-air meteorological situation (temperature, wind, airmass convective stability, sunshine, and precipitation). Others included are persistence (V01–V06), sea-breeze development potential (V25–V32), and aging of the summer season (V41). The same predictors are not always chosen by CART at each site, particularly sites far removed from each other. Several of the predictors are correlated, but as stated in the appendix, parsimony among predictors is not a problem. This is a particular advantage for groups of the same type of predictor for different sites, for example, V01–V06, where some observations, but not all, can be missing on a particular day.

Summary results of the CART analyses for all regions are plotted in Fig. 3. In Fig. 3a, we see that the variance in the data for three of the Vancouver region sites is the lowest anywhere. Figure 3b shows that CART developed trees with approximately 20 nodes for the Vancouver sites, except for the 11-node tree for Robson Square. According to the cross-validation (CV) error and resubstitution error estimates, we see in Fig. 3c that by fitting the data with CART trees the variances are considerably lowered relative to the initial value for all the regions, with inland sites in the Vancouver region showing the best improvement. For the Vancouver region, the unexplained variance implied by the resubstitution error ranges from 43%–55% at the urban sites, which have very low initial variance, to 24%–31% at the inland sites, where initial variance is much higher. The CV errors are higher but, as previously mentioned, give a more realistic estimate of the unexplained variance to be expected when using the trees for prediction. For example, at Rocky Point Park the resubstitution variance estimate of the error is 0.24 $\times$ 326 $(\text{ppb})^2$ (standard deviation of 8.8 ppb), but an unbiased estimate of the error to be expected in predictions is given by the CV error of 0.43 $\times$ 326 $(\text{ppb})^2$, which is 140 $(\text{ppb})^2$ (standard deviation of 11.8 ppb). It is not known how much of this remaining variance can actually be explained and how much is "noise" due to variability on time and space scales too small to model with the predictors used.

CART's ranking of predictor importance (not shown) for inland sites is usually different from the two urban sites, reflecting differences in the climate. The overall most important group of predictors is the surface maximum temperature or its anomaly (V07–V18). Hours of sunshine (V40) is the most

TABLE 1. Predictors for Vancouver sites. Locations of O_3 and meteorological observation sites are shown in Fig. 1, except Tofino is about 175 km west of Vancouver and Penticton is about 225 km east of Vancouver. Upper-air meteorological sounding data taken at 1200 UTC is from Quillayute, Washington, (UIL), about 150 km southwest of Vancouver. Precipitation occurrence is defined as 0 if none occurred, 1 if occurred. Dry convective mixing height in V33–V38 is found by lifting an air parcel at the daily maximum temperature for the specified site dry adiabatically from the surface until the UIL sounding temperature is reached. The denominator in V41 is 366 in leap years. Units: pressure (hPa), temperature (°C), relative humidity (%), wind speed (kt), wind direction (°), geopotential height (dam), height of maximum potential temperature (m), sunshine (h), O_3 (ppb), dry convective mixing height (m).

	Predictor		Predictor
V01	Robson Square maximum O_3 yesterday	V30	Vancouver–Abbotsford pressure difference 0000 UTC
V02	Richmond South maximum O_3 yesterday	V31	Vancouver–Hope pressure difference 0000 UTC
V03	Rocky Point Park maximum O_3 yesterday	V32	Vancouver–Penticton pressure difference 0000 UTC
V04	Surrey East maximum O_3 yesterday	V33	Vancouver dry convective mixing height
V05	Abbotsford maximum O_3 yesterday	V34	Haney dry convective mixing height
V06	Chilliwack maximum O_3 yesterday	V35	Abbotsford dry convective mixing height
V07	Vancouver maximum temperature	V36	Agassiz dry convective mixing height
V08	Haney maximum temperature	V37	Chilliwack dry convective mixing height
V09	Abbotsford maximum temperature	V38	Hope dry convective mixing height
V10	Chilliwack maximum temperature	V39	Abbotsford surface wind speed 1800 UTC
V11	Agassiz maximum temperature	V40	Abbotsford sunshine
V12	Hope maximum temperaure	V41	$10^4 \sin[(\text{Julian day} - \text{Julian day March 21})2\pi/365]$
V13	Vancouver maximum temperature anomaly	V42	UIL sounding potential temperature difference (850–1000 hPa)
V14	Haney maximum temperature anomaly	V43	UIL sounding maximum temperature
V15	Abbotsford maximum temperature anomaly	V44	height of UIL maximum temperature
V16	Chilliwack maximum temperature anomaly	V45	UIL maximum potential temperature (surface to 500 hPa)
V17	Agassiz maximum temperature anomaly	V46	height of maximum potential temperature in V45
V18	Hope maximum temperature anomaly	V47	850-hPa temperature
V19	Vancouver precipitation	V48	850-hPa wind direction
V20	Haney precipitation	V49	850-hPa wind speed
V21	Abbotsford precipitation	V50	850-hPa geopotential height
V22	Chilliwack precipitation	V51	700-hPa temperature
V23	Agassiz precipitation	V52	700-hPa wind direction
V24	Hope precipitation	V53	700-hPa wind speed
V25	Vancouver–Tofino pressure difference 1200 UTC	V54	700-hPa geopotential height
V26	Vancouver–Abbotsford pressure difference 1200 UTC	V55	500-hPa wind direction
V27	Vancouver–Hope pressure difference 1200 UTC	V56	500-hPa wind speed
V28	Vancouver–Penticton pressure difference 1200 UTC	V57	500–1000-hPa geopotential height difference
V29	Vancouver–Tofino pressure difference 0000 UTC		

important predictor of all at either end of the valley (Robson Square and Chilliwack) and is also very important elsewhere. The importance of temperature and solar radiation predictors is to be expected since the photochemical reactions that produce O_3 are directly related to them (see McKendry 1993). The strong correlation between maximum temperature and surface O_3 has also been noted by Robeson and Steyn (1990), Stockenius (1991), McKendry (1993), and many others. Persistence predictors (V01–V06) are very important at the station of measurement; for example, V06 is very important at Chilliwack but is less important elsewhere. The next most important group of predictors overall are those defining or related to airmass temperature and convective stability, in particular V42–V47, V50, V51, V54, V57. The more specific measure of convective stability (V33–V38) is slightly outranked by less specific measures (V42–V47), particularly at the inland stations. Surface pressure gradient predictors V27–V29 are predictors of low-intermediate importance, more so for the urban sites. Precipitation predictors are much more important at the two urban sites than inland. Predictors at 700 and 500 hPa were not important in an overall sense, but they did occasionally show up in node-splitting rules involving occurrences of high O_3, as will be seen below.

Figure 4 shows the decision-tree information for Rocky Point Park. The architecture of the tree is Fig. 4a. Data enter the tree at the top and drop down through a series of internal nodes until a terminal node is reached. The path to reach terminal node 6, for example, would be a branch left in internal node 1, right in internal node 2, left in internal node 6, and right in internal node 7. In Fig. 4b, we see that generally higher O_3 values have been separated on the right side of the tree in Fig. 4a from generally low values on the left side. The standard deviations of the cases in the terminal nodes rises from about 7–10 ppb in nodes 1–8 to about 10–15 ppb in nodes 9–21. There are far fewer cases of high O_3 concentration than low concentration, as seen from Figs. 4b and 4d. In Fig. 4c we see that terminal nodes 12–13 and 16–20 have nonzero probabilities of O_3 concentration of at least 82 ppb.

Splitting rules for internal nodes of the tree in Fig. 4 appear in Table 2. Examination shows them to make

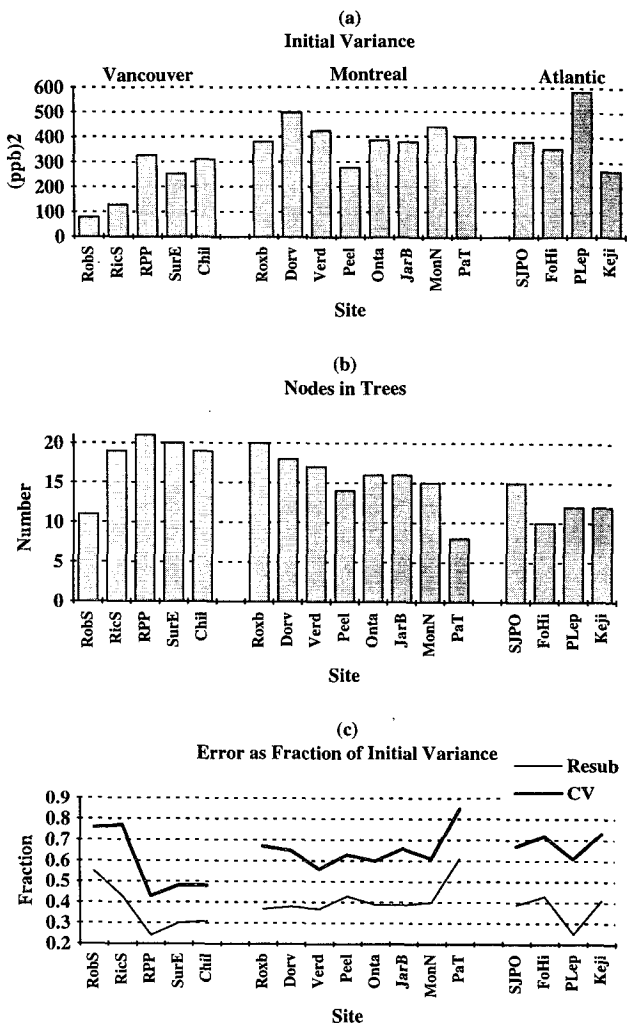

FIG. 3. Summary statistics for CART analyses. Groups of sites and abbreviations of site names as in Fig. 2. Graphs: (a) initial variance in observed maximum O_3 data [(ppb)^2]; (b) number of nodes in decision trees; (c) resubstitution error (resub) and cross-validation error (CV), expressed as a fraction of initial sample variance.

sense physically. Occurrence of maximum O_3 concentrations of at least 82 ppb at Rocky Point Park can be determined by maximum air temperature (V09), maximum O_3 the previous day (V03 and V04), Julian day (V41), sea-breeze strength in the morning (V27 and V28) and afternoon (V31), and 700-hPa wind direction (V52). The rule for internal node 1 shows that cases where O_3 is generally high can be separated from those where O_3 is low if the Abbotsford maximum temperature (V09) is greater than 24.7°C. Chances for O_3 concentration exceeding 82 ppb are greatest when conditions for branching to terminal nodes 16, 18, 19, or 20 occur. First, the Abbotsford maximum temperature must be greater than 29.8°C, as seen by the rule for internal node 9. Branching to terminal nodes 16 and 17 occurs via internal node 16 if the maximum O_3 at Surrey East (V04) is less than or equal to 60.2 ppb the previous day. According to the rule for internal node 17, if the 700-hPa wind direction is less than 133° (i.e., east-southeast through north-northeast), a case goes to terminal node 16 and is assigned an O_3 level of 79 ± 14 ppb and probability 0.43 of $O_3 \geqslant 82$ ppb. Otherwise the case goes to terminal node 17 and is assigned a lower O_3 concentration 61 ± 15 ppb and reduced probability 0.12 of $O_3 \geqslant 82$ ppb. If V04 > 60.2 ppb, then terminal nodes 18–21 are reached. For these cases, if V41 ⩽ 4010 (date is 27 August or later), then terminal node 18, with maximum O_3 average value 132 ± 11 ppb, is reached. Otherwise, if the

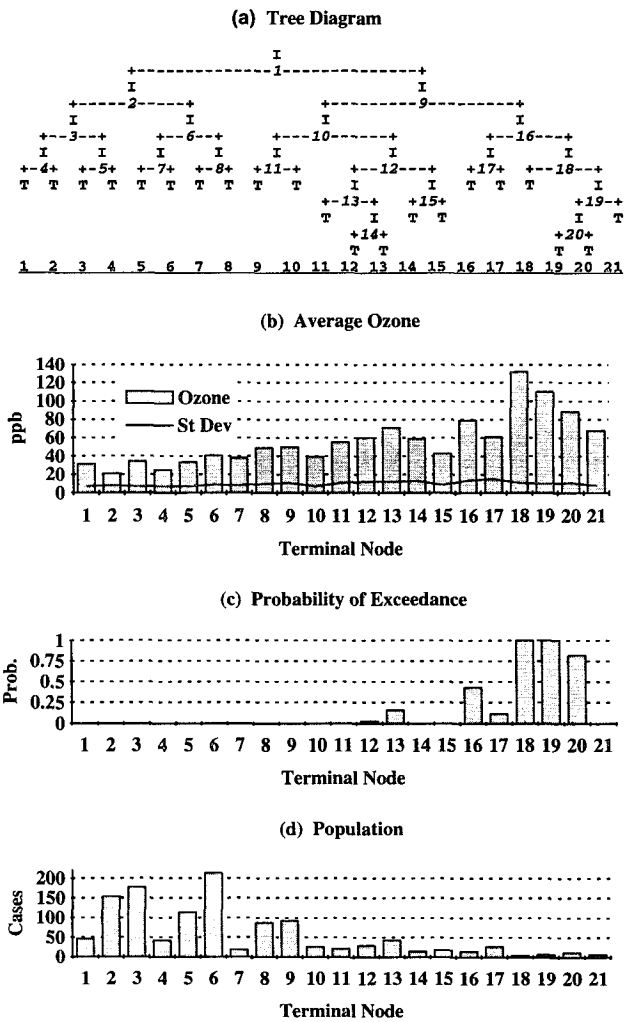

FIG. 4. CART tree information from learning data for Rocky Point Park. Graph (a) is a tree diagram: Data enters tree at the top and drops down through internal nodes until a terminal node is reached. Internal nodes are numbered in italics at node locations and marked with "I" above. Terminal nodes are marked with a "T" below the node location and numbered directly below; (b) average O_3 value and standard deviation of cases in terminal nodes (ppb); (c) fraction of cases with observed maximum $O_3 \geqslant 82$ ppb in terminal nodes (probability of exceedance); (d) number of cases in terminal nodes.

TABLE 2. Splitting rules for internal nodes in Rocky Point Park tree in Fig. 4. A case branches left from an internal node if the rule calculated from the predictors in column 2 is less than or equal to the threshold value in column 3.

Node	Rule	Threshold value
1	V09	24.7
2	V14	−1.85
3	V40	1.35
4	−0.710 (V13) −0.263 (V32)	7.93
5	−0.486 (V27) −0.300 (V30) −0.0611 (V02) −0.310 (V40) −0.052 (V14) −0.352 (V19) + 0.00744 (V28) + 0.0916 (V45) + 0.658 (V21)	−0.704
6	V16	3.05
7	0.285 (V01) +0.00111 (V41) −0.467 (V39) −0.836 (V29)	6.59
8	V03	31.1
9	V09	29.8
10	V03	46.9
11	0.389 (V30) +0.114 (V12) + 0.912 (V21) −0.0471 (V01) −0.000969 (V48)	2.63
12	V28	2.55
13	V31	0.45
14	V03	57.1
15	V39	13.9
16	V04	60.2
17	V52	133
18	V41	4010
19	V27	−0.850
20	V31	1.05

date is earlier than 27 August and V27 ≤ −0.85 hPa (pressure increasing eastward along the Fraser River valley in the morning, which would retard the sea breeze), then terminal nodes 19 or 20 are reached. The higher-valued terminal node 19 is reached if the afternoon pressure gradient along the valley is still easterly or is weak westerly (V31 ≤ 1.05 hPa). Figures 4a,c show there is a small chance of an O_3 exceedance of at least 82 ppb if a case branches left from internal node 9 to terminal nodes 12 or 13. This can occur if the Abbotsford maximum temperature is between 24.7° and 29.8°C, at least 46.9-ppb maximum O_3 concentration was measured the day before at Rocky Point Park, there was less than 2.55-hPa pressure difference at 1200 UTC from Vancouver to Penticton, and the Vancouver to Hope pressure difference at 0000 UTC was less than 0.45 hPa.

b. Montreal region

The city of Montreal lies primarily on a large island complex and adjacent shore area in the St. Lawrence River valley. The valley runs southwest to northeast and is about 120 km wide in the Montreal area. Maximum surface O_3 concentrations were analyzed for eight sites shown in Fig. 5. All are in the urban area of Montreal, none are adjacent to major highways. Roxboro and Dorval are in western residential areas, Verdun in a residential area just south of the downtown area, Peel is in the heart of the downtown core, Ontario is on the northeast periphery of the downtown area, Jardin Botanique is in the residential east city, Montreal Nord is in the residential northeast end of the city, and Pointe aux Trembles is in the northeast part of the city just east of an oil refinery complex. Several previous studies (e.g., Chung 1977; Mukammal et al. 1982; Heidorn and Yap 1986) found this region of eastern Canada to be characterized by the long-range transport of O_3 and its precursors northeastward along the heavily populated corridor from southern Lake Erie up the St. Lawrence River valley. High O_3 concentrations are most often seen in summer during times of high temperature and west to southwest winds, conditions that occur when the western flank of a large anticyclone lies over the area. The surface wind is channeled by the river valley, preferred directions are from the west and southwest.

Data were available for the Montreal region at most stations continuously since 1980, the longest period of the three regions analyzed. The population of the Montreal area has not changed substantially, and no trend of O_3 concentration has been detected during this time. Figure 3a shows the average O_3 concentration is highest at the western sites, decreases toward the downtown core, reaching a minimum there, then slowly increases northeastward with distance from downtown. This pattern was also noted by McKendry (1993) and is consistent with the concept of air with relatively high O_3 content impinging on the upwind (western) side of the city being scavenged by nitrogen oxides emitted from vehicular traffic and industry in the urban environment, then a gradual increase of O_3 concentration downwind of the urban core. Here, as

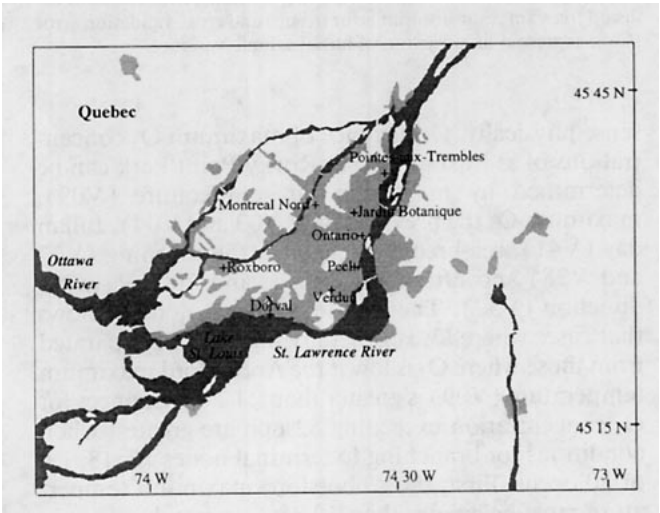

FIG. 5. Montreal region. Air monitoring sites marked with "+"; surface climate data site with "×." Urban areas shaded lighter than water areas.

elsewhere in Canada, O_3 concentration exceeding 80–82 ppb is a relatively rare event. Figure 2b shows the probability of maximum daily O_3 concentration of at least 80–82 ppb is about 0.05–0.08 on the west side of Montreal, decreases to near 0 toward the city center, then slowly increases with distance east and north of the city center to about 0.03 at the northeast end of the city.

Predictors used for the CART analysis are given in Table 3. Upper-air meteorological predictors are from the Maniwaki radiosonde station, about 180 km northwest of Montreal. Surface meteorological predictors are measured at Dorval airport, which is local. Many predictors are the same type as for Vancouver, with the additions of dominant wind, surface dewpoint, haze, and synoptic class. The previous day's O_3 observation is for the specific site being analyzed rather than all sites. Maximum temperature at Dorval airport is used with Maniwaki sounding data to compute the dry convective mixing height. The dominant wind speed and direction, defined as the most frequently observed wind speed and direction during the 24-h period when the O_3 maximum occurred, are used as surrogates for long-range transport of O_3 and its precursors. Specific long-range transport predictors such as back-trajectory locations and upwind O_3 will be included in a future predictor set.

Error statistics for the CART analyses of the Montreal sites are shown in Fig. 3. Initial variances in the maximum O_3 data are higher than for the Vancouver region, except at the Peel site, which has much lower O_3 levels than elsewhere in the Montreal region. The higher variance is likely a result of more complicated meteorology and transport of O_3 and its precursors from large upstream urban areas. There are approximately 15–20 nodes in the trees, except for 8 at the northernmost site (Pointe aux Trembles). The data for that site stands out in Fig. 3c as more poorly fit by CART than the others. Whether this is due to predictors that should be included but are not, or that the chemistry is more complex at this station due to its proximity to an oil refinery complex, is not known. At the remaining sites the resubstitution errors are 0.37–0.43, implying the variance not explained in the learning data is about 40%. The relative CV errors at these sites are 0.56–0.67, thus a significant reduction in the error is achieved by using the trees to predict O_3. While this error level is good, it is higher than for Vancouver, thus there likely is room to improve expected prediction skill by adding predictors more specifically related to pollution transport from upstream areas.

Figure 6 shows the CART ranking of predictor importance. The order of predictors plotted was determined by descending order of predictor importance for Verdun. There is an envelope of decreasing rank, with some variation in the order of the five most important predictors between the two western sites and the others. The similarity of predictor importance among sites is not surprising since they are in a single urban area, with the same meteorological factors controlling O_3 concentrations at all sites. The most important predictors besides persistence are those defining the temperature and humidity of the air mass over the city. Most important overall are surface maximum temperature (M02) and the previous day's O_3 maximum (M01). Four major secondary predictors are seen: 850-hPa temperature (M19), 500–1000-hPa thickness (M24), surface dewpoint (M03), and 850-hPa geopotential height (M22) at the northernmost site. There is a group of predictors of intermediate importance: haze (M14), Julian day (M17), afternoon surface wind direction (M06), dominant wind direction (M09), surface relative humidity (M04), sunshine (M08), and surface–850-hPa temperature difference (M23). The remaining predictors are of low importance but still ranked 10–20 at many sites. Four predictors are ranked well below 10 at all sites: day of the week (M16) and the precipitation predictors M11, M12, and M13.

As noted above, surface dewpoint is an important secondary predictor. Lelieveld and Crutzen (1990) noted that photochemical models that do not consider the role of clouds will considerably overestimate ozone

TABLE 3. Predictors for Montreal sites. Surface meteorological data are at 1800 UTC from Dorval airport. Upper-air meteorological data are at 1200 UTC from Maniwaki radiosonde station, about 180 km northwest of Montreal. Surface relative humidity is calculated from maximum temperature and dewpoint at 1800 UTC. Synoptic classes are from Heidorn and Yap (1986). Precipitation, thunderstorm, and haze occurrences are defined as 0 if none occurred, 1 if occurred. The day of the week is numbered 1–7 starting with Sunday. Units: temperature (°C), relative humidity (%), wind speed (kt), wind direction (°), geopotential height (dam), sunshine (h), O_3 (ppb).

	Predictor
M01	maximum O_3 yesterday
M02	maximum temperature
M03	surface dewpoint
M04	relative humidity from M02 and M03
M05	surface wind speed
M06	surface wind direction
M07	surface pressure
M08	sunshine
M09	dominant wind direction
M10	dominant wind speed
M11	24-h precipitation
M12	precipitation 1200–1800 UTC
M13	thunderstorm
M14	haze
M15	synoptic class (1–8)
M16	day of the week
M17	Julian day of the year
M18	Dorval dry convective mixing height
M19	850-hPa temperature
M20	850-hPa wind direction
M21	850-hPa wind speed
M22	850-hPa geopotential height
M23	surface–850-hPa temperature difference
M24	500–1000-hPa geopotential height difference

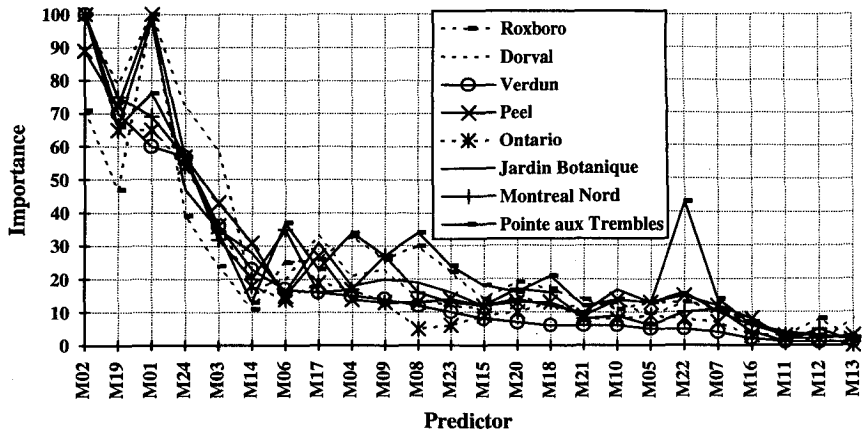

FIG. 6. Importance ranking of predictors for Montreal region.

concentrations as sunlight is scattered and reflected by clouds and important chemical reactions occur. When specific decision trees were inspected, dewpoint and the correlated predictor, relative humidity, were found to have a limiting role in episodes of high ozone concentration. For example, in the tree for Verdun (not shown), factors favoring high ozone concentration (high temperature, low wind speed, and low mixing height) led to a pool of 51 cases with average O_3 value of 88 ppb in one node. From there CART split off a group of 38 cases whose average ozone was 81 ppb based on dewpoint greater than about 19°C (the first competitor rule was relative humidity greater than about 41%), while average O_3 for the remaining 13 cases was 108 ppb.

c. *Atlantic region*

Figure 7 shows the air monitoring and climate data sites, which are scattered over a much larger area than the Vancouver and Montreal sites. Forest Hills and Saint John airport are in the small city of Saint John, Point Lepreau is on a point of land jutting into the Bay of Fundy, and Kejimkujik is in an isolated national park in Nova Scotia. Analysis was done for a fifth site, Norton, but results are not shown since only two years of O_3 measurements were available.

Statistics for maximum O_3 concentrations are in Fig. 2. Only five years of data, 1986–90, were available for study at Point Lepreau, and six years, 1985–90, were available for the other sites. Average O_3 levels are comparable with the Vancouver inland sites and the Montreal sites (except for Peel). The probability of maximum O_3 concentration of at least 82 ppb is about 0.02–0.03 at all the sites except Point Lepreau, where it is over 0.06, but there were only half as many observations there as for the other sites. There are no large urban areas here; however, when warm stable air masses are present, the region is subject to high O_3 levels due to long-range transport of pollutants produced in large cities to the southwest along the Atlantic coastal area of the United States and the southern Great Lakes–St. Lawrence River corridor. The higher exceedance numbers at Point Lepreau may be due to pollutants transported over water. Pollution of relatively high concentration can travel great distances over cold coastal waters because it becomes trapped at the top of a shallow, stable marine layer, resulting in lower deposition rates than over land. About half the O_3 exceedances of at least 82 ppb occur at night in the Atlantic region, and exceedances can occur in cloudy or sunny weather. Since there are more complicating factors here than elsewhere in Canada, prediction of high O_3 concentrations is expected to be more difficult.

FIG. 7. Atlantic region. Air monitoring sites marked with "+"; climate data sites marked with "×." Water areas are shaded.

The surface meteorological predictors used for each site are similar to those used for Montreal, except they are at the hour of O_3 maximum instead of midafternoon. Surface data are measured at Saint John airport for Saint John Post Office, Forest Hills, and Point Lepreau, at Fredericton airport for Norton, and at Greenwood airport for Kejimkujik Park. Maximum temperature, precipitation, haze, and sunshine are from the previous day if the O_3 maximum occurred between midnight and 0600 local time. Wet-bulb temperature was used in place of dewpoint, but they are similar measures of airmass moisture content. Cloud opacity and amount were added as predictors; thunderstorm occurrence, dominant wind speed and direction, and day of the week were dropped. The hour of occurrence was included as a predictor because a significant proportion of ozone maximum concentrations occur at times other than midafternoon. The upper-air predictors are the same as used for Vancouver–Lower Fraser Valley and are taken from the 1200 UTC sounding at Caribou, Maine, about 200 km north of St. John. Specific long-range transport predictors such as back-trajectory locations and upwind O_3 are not included here but will be in a future predictor set.

The rankings of predictor importance showed there is one dominant predictor, the O_3 maximum the day before, at three of the sites. At Forest Hills, a second predictor, the 850–1000-hPa potential temperature difference, is also very important. Predictor rankings for Kejimkujik Park are different from the other sites and are similar to the Montreal sites, probably because it is inland whereas the others are near the sea. Haze is of intermediate importance at all sites. At various sites near the Bay of Fundy, sounding maximum temperature, mixing height, wind direction, and Julian day predictors are of intermediate importance. The only predictor of importance less than 10 at all stations is precipitation occurrence. All others have low but significant importance at most sites.

Error statistics for the CART analyses of the Atlantic sites are in Fig. 3. Initial variances in the O_3 data are comparable to Montreal at the two St. John sites but are higher at Point Lepreau and lower at Kejimkujik. There are 10–15 nodes for the CART trees, fewest of the three regions studied, an indication that fitting the data was more difficult. The resubstitution errors imply the variance not explained in the learning data is about 25% at Point Lepreau and about 40% at the other sites. The CV errors of 0.61–0.73 are comparable to the Montreal region but are higher than the Vancouver inland sites. Thus, the CART trees are expected to have predictive skill but there could be room for improvement, possibly by the addition of predictors specifically related to pollution transport from upwind urban areas and sea-breeze development potential at coastal sites.

3. Estimating performance with independent data

a. Method

The question arises as to how the CART decision-tree analyses will perform when used to make forecasts of O_3 exceedances with independent data. The CART trees are already selected using cross-validation testing to estimate their error when used with independent data in order to find an honest tree architecture. However, error in forecasts of the predictors is not considered. A full test entails testing the trees independently in the field with forecaster input of the predictors. Due to large interannual variability in the number of days where O_3 concentrations exceed 80 ppb, in Canada it is desirable to test the forecast procedure with several years of data. It is not feasible to wait for several more years, so a substitute method for testing was used that, though not as satisfactory as a proper test, we believe to be reasonable for giving an expectation of performance of the trees with independent data. The test consisted of withholding one year of data, growing CART trees with the remaining data using the same complexity parameter as the trees grown with the full dataset (this gives approximately the same number of nodes), testing them with data withheld, then repeating with a different year withheld for each of the five years 1988–92. Tests were done for the Vancouver and Montreal regions but not for the Atlantic region since fewer years of data were available there. Because maximum O_3 was rounded off to the nearest 10 ppb for Montreal, an exceedance will be defined for test purposes here as $O_3 \geqslant 80$ ppb for both regions.

In practice, the forecasts for the afternoon of the current day would be generated the previous day and updated the morning of the current day when the 1200 UTC upper-air data becomes available (1200 UTC is 0800 LST in the Atlantic region, 0700 LST in the Montreal region, and 0400 LST in the Vancouver region). The test still assumes that correct values of all predictors are available. The errors reported here should be about the same as for the updated forecasts of the current day since many predictors would be known and only a few would have to be forecast for the afternoon. Greater error would normally be expected in the forecasts prepared the previous day since many predictors would have to be forecast about 18–30 h in advance. However, the discrepancy will hopefully not be serious in most cases given the high accuracy of modern NWP model forecasts in this time range. This would be especially true for situations where forecast predictor values are not close to decision threshold values in internal nodes, so that even significant differences in forecasts of predictors made the previous day and the current morning would not change the forecast.

Ozone concentrations exceeding 80 ppb are relatively rare events in Canada. A forecaster will have available to him a list of forecasts for all sites in his region. As a strategy to prepare forecasts for a small area that includes several sites, it is often advantageous to use the ensemble of forecasts. In fact, this is usually done anyway by the forecaster since the area of responsibility is generally an urban area rather than a specific location. Since any statistical method becomes

TABLE 4. Definition and description of CART forecast and observation classes for ensembles of sites at Montreal and Vancouver. Type 1 forecast is CART node average of at least 80 ppb. Type 2 forecast is CART node average less than 80 ppb but node average plus one-half standard deviation is at least 80 ppb and probability of exceedance is at least 0.2. Written numbers pertain to number of sites (e.g., ⩾ two type 1 means type 1 forecasts at two or more sites).

Forecast class	Description	Definition	Observation class	Description	Definition
Montreal					
F1	exceedance alert	⩾ two type 1 or one type 1 + ⩾ two type 2 or ⩾ 3 type 2	OB1	exceedance	⩾ two sites ⩾ 80 ppb
F2	exceedance check	one type 1 or two type 2	OB2	near exceedance	one site ⩾ 80 ppb
F3	elevated O_3	one type 2 or maximum O_3 60–79 ppb	OB3	elevated O_3	maximum O_3 60–79 ppb
F4	low O_3	maximum $O_3 < 60$ ppb	OB4	low O_3	maximum $O_3 < 60$ ppb
Vancouver					
F1	exceedance alert	⩾ one type 1 or ⩾ two type 2	OB1	exceedance	⩾ one site ⩾ 80 ppb
F2	exceedance check	one type 2	OB2	elevated O_3	maximum O_3 60–79 ppb
F3	elevated O_3	maximum O_3 60–79 ppb	OB3	low O_3	maximum $O_3 < 60$ ppb
F4	low O_3	maximum $O_3 < 60$ ppb			

less reliable when forecasting rare events, we expect a higher level of error when forecasting for single sites than for small areas. As noted in Burrows (1991), it was often seen that a rare event that occurred at a site may not have been forecast to occur at that particular site but was forecast to occur at some nearby site(s) within a small region. If the rare event did not occur at the nearby site(s), single-site verification could lead one to conclude that the forecasts were wrong at all sites, when in fact the rare event had actually been predicted to occur within the small region. This situation can occur with CART forecasts when a particular decision rule in the later internal nodes causes an event to branch in the wrong direction because the value of the rule calculated for the event is near the threshold value but on the wrong side of it. Even if this does occur, when the predictors are known to bear good physical relation to the response variable (as is the case here) the forecasts will often be "in the ballpark" when they are wrong. The applicable areas for public forecasts of O_3 exceedances in Canada are mainly small areas with several sites, so it is reasonable to take the ensemble approach for prediction and verification here.

For O_3 exceedances there are two types of CART forecasts to consider. A "type 1" forecast occurs when a terminal node, whose average O_3 concentration is at least 80 ppb, is reached. A "type 2" forecast occurs when a terminal node is reached whose average O_3 concentration is below exceedance level but there is a significant probability of exceedance associated with the node. After some experimentation, the following system was adopted for exceedance prediction in type 2 situations: if the node average plus one-half the standard deviation is at least 80 ppb and the exceedance probability is at least 0.20, consider the forecast as an exceedance forecast but with about one-half the weight of a type 1 forecast.

In designing a method to use and verify an ensemble of CART forecasts for O_3 exceedances, we need to define when an exceedance is observed and when it is forecast. For Vancouver there is a realistic chance of an exceedance only at the three inland sites, and for Montreal at seven of the eight sites. An exceedance for Vancouver was deemed to be observed if at least one of the sites reported at least 80 ppb, and for Montreal if at least two of the sites reported at least 80 ppb. To recognize near misses and "right idea" forecasts, the hierarchy of types of forecasts and observation classes shown in Table 4 was designed for the Montreal and Vancouver ensembles. The most important forecast and observation classes are the F1 and OB1 designations, respectively.

b. Results

Results of the tests are shown in Table 5 for each city and year. The left section of the table shows totals for OB1 observed exceedances and matching F1–F4 forecasts on the days of those exceedances. The right section shows totals for F1 forecast exceedances and matching OB1–OB4 observations. Large interannual variability in observed exceedances is seen and is attributable to year-to-year differences in meteorological conditions, a normal occurrence in Canada. In general, the number of exceedances forecast is about the same as the number observed, thus the forecast method can be expected to behave reasonably as meteorological conditions fluctuate from one year to the next.

Considering the probability of occurrence of an exceedance is at best only a few percent at any site, we believe the CART forecasts show reasonably good skill overall. On the days of OB1 observed exceedances at Montreal, approximately 41% were correctly forecast with F1 forecasts, and there were F1 or F2 forecasts for approximately 61% of them. Similar figures for Vancouver were about 47% and 65%, respectively. For Montreal there were F1, F2, or F3 forecasts on about 88% of the days on which OB1 exceedances were observed, and this figure for Vancouver was about 79%. On the days of the F1 forecasts for Montreal, there

TABLE 5. Results of test for estimating decision-tree performance with independent data. Definitions are in Table 4.

Year	OB1	F1	F2	F3	F4	F1	OB1	OB2	OB3	OB4
Montreal										
1988	23	11	5	5	2	15	11	3	1	0
1989	15	4	4	5	2	17	4	3	8	2
1990	5	1	1	2	1	2	1	0	1	0
1991	9	4	1	3	1	7	4	0	1	2
1992	4	3	0	0	1	6	3	2	1	0
Total	56	23	11	15	7	47	23	8	12	4
Fraction	1.00	0.411	0.196	0.268	0.125	1.00	0.489	0.170	0.255	0.085
Vancouver										
1988	19	8	5	3	3	9	8	1	0	
1989	5	2	0	3	0	7	2	2	3	
1990	12	10	1	0	1	12	10	1	1	
1991	0	0	0	0	0	2	0	1	1	
1992	7	0	2	0	5	4	0	2	2	
Total	43	20	8	6	9	34	20	7	7	
Fraction	1.00	0.465	0.186	0.140	0.209	1.00	0.588	0.206	0.206	

were O_3 exceedances observed at two or more sites for approximately 49%, and at one or more sites for about 66%. About 92% of the F1 forecasts for Montreal were matched with $O_3 \geq 80$ ppb or elevated O_3 of 60–79 ppb at one or more sites. On the days of the F1 forecasts for Vancouver, about 59% were matched with observed O_3 exceedances at one or more sites, and about 79% were matched with exceedances or elevated O_3 at one or more sites.

The degree of year-to-year homogeneity in accuracy of forecasts is hard to quantify with a statistic because we are looking at forecasts of an extreme event with relatively few occurrences and large differences in the number of events between years. The Montreal results should be more representative because there are seven "exceedable" sites and only three for Vancouver. Looking at the year-by-year OB1 exceedances, about 0.4–0.75 were correctly forecast with either F1 or F2 forecasts for Montreal, as well as for Vancouver in three of the five years. For the F1 forecasts, for Montreal about 0.5–0.75 coincided with OB1 observations (except for 1989) and about 0.5 to greater than 0.75 coincided with either OB1 or OB2 observations, while about the same is true for Vancouver OB1 exceedances in four of the five years (counting 1991 as a success since no exceedances were observed and only two were forecast). The number of "bust" forecasts (F4 O_3 forecast when OB1 occurred and the converse) is generally in the range 10%–20% for both regions in all years.

Remembering that the forecasts would be used as guidance by a forecaster, several of the F2 or even F3 forecasts would possibly have been upgraded to an exceedance alert, depending on his judgment of the meteorological situation. It is expected that with use of the forecasts over time a methodology for interpretation and adjustment of the forecasts would be developed by the users.

4. Summary and conclusions

In the spring of 1993 Environment Canada began daily preparation of air quality forecasts for the Vancouver–lower Fraser River valley, Montreal, and Atlantic regions of southern Canada where air quality has become a concern in recent years. An advisory is issued jointly with provincial authorities if the 1-h maximum surface O_3 concentration is expected to exceed 82 ppb. Occurrences of high O_3 concentrations are episodic and rather rare in Canada. The relation between surface O_3 and meteorological predictors is nonlinear, thus nonparametric decision-tree regression models are appropriate. The purpose of this paper is to report on the development and testing of site-specific statistical models for analysis and forecasting of surface ozone, where a multivariate data-driven rule-based nonparametric statistical analysis procedure known as CART was used.

CART decision-tree analysis was done for five sites in the Vancouver–lower Fraser River valley region, eight sites in the Montreal urban area, and four sites in the Atlantic region. Surface and upper-air meteorological predictors and other predictors were matched with observed daily maximum O_3 concentrations for the months of May–September for several years. The decision trees were found to fit the data reasonably well, and the rules for node splitting were found to be physically realistic. A brief description of the CART procedure is given and some of the results of the analyses for the three regions are discussed, with emphasis on factors leading to high O_3 concentrations. The most important predictors besides persistence were those defining airmass temperature, convective stability, moisture content, and sunshine. A significant finding for the Montreal region was that even with high temperatures and southwest winds, elevated moisture content in an air mass (dewpoint about 19°C or higher)

will decrease O_3 concentration. This may be a consequence of reduced photochemical reaction rates due to increased scattering of sunlight and important chemical reactions. Airmass moisture content was also an important predictor in the Atlantic region but was not used in the Vancouver region. The variance explained in the Vancouver region was higher than the Montreal and Atlantic regions. This is likely a consequence of more variance in the training data in the latter two regions and may be an indication of the more complicated meteorology there and a need for more specific predictors related to long-range transport of O_3 and its precursors, such as back-trajectory locations and upstream O_3 values.

The CART trees are already selected using cross validation to estimate their error when used with independent data in order to find an honest tree architecture. A method for estimating performance of the decision-tree models with independent data for forecasts of O_3 concentrations exceeding 80 ppb in individual years was tested for Vancouver and Montreal. The method follows the strategy a forecaster is likely to follow when issuing public forecasts valid for a small region where several monitoring sites are located, by using the ensemble of separate forecasts to make one forecast that applies everywhere in the region as opposed to using the forecasts site by site. The forecasts show good skill considering that occurrences of O_3 concentrations exceeding 80–82 ppb are relatively rare in Canada. About 66% of the forecasts of maximum daily O_3 concentration of at least 80 ppb at two or more sites for Montreal were matched with observed $O_3 \geq 80$ ppb at one or more sites, and about 92% were matched with observed $O_3 \geq 80$ ppb or elevated O_3 of 60–79 ppb at one or more sites. About 59% of the forecasts of $O_3 \geq 80$ ppb at one or more sites for Vancouver were matched with observed $O_3 \geq 80$ ppb at one or more sites, and about 79% were matched with observed $O_3 \geq 80$ ppb or elevated O_3 concentrations of 60–79 ppb.

We believe the method outlined here for surface O_3 prediction is capable of producing reliable forecasts for operational use in issuing public forecasts. A computer program, which does the calculations upon input of predictors, some by the user and some by automatic injection of NWP model output data from the Canadian Meteorological Centre, has been provided to each of the three regions. The decision trees can be easily updated in subsequent years as data become available. It is expected that with use of the forecasts over time, a methodology for interpretation and adjustment of the forecasts would be developed by the users that would further increase forecast accuracy.

New decision trees are planned for the future. Larger predictor sets, which include back trajectories and upstream O_3 and other air pollutants, are being developed to better handle the long-range transport process. Decision trees are planned for additional sites and possibly for other pollutants. For upper-air predictors, since the radiosonde stations are not locally located in any of the regions, we are considering replacing radiosonde data with objective analysis data at the site locations.

Acknowledgments. We would like to express our appreciation to the provinces of British Columbia, Quebec, New Brunswick, and Nova Scotia for collecting and providing the O_3 data. We are grateful to Claude Gagnon of the Environment Service, Communauté Urbaine de Montréal, for providing O_3 data and for his expertise on local O_3 behavior. Thanks are due to Keith Keddy of Environment Canada Atlantic Region for compiling the data for the Atlantic sites and to Walter Gilles of Environment Canada Pacific Region for compiling the data for the Vancouver–lower Fraser River valley sites. We wish to thank Charles L. K. Paterson for designing, writing, and implementing the computer programs provided to the regional forecast offices. We also acknowledge Dr. S. Venkatesh, Dr. A. McMillan, Dr. S. C. Pryor, and the anonymous reviewers for their valuable reviews of the original manuscript.

APPENDIX

CART

Since CART is not yet widely known in the meteorological community, a few comments follow on the CART regression procedure and some features that experience has shown are useful for data analysis. Detail can be found in Breiman et al. (1984).

CART is essentially a top-down data-driven procedure where rules and a rule architecture that fit the predictand data as well as possible are developed systematically from the predictors. It is different from a bottom-up tree-based procedure where a predetermined rule architecture is forced by successive iterations to fit new data, such as in a back-propagation neural net. From a learning database of matched predictand and user-defined predictors, CART develops a treelike structure of decision nodes to split high-valued cases from low-valued cases by minimizing an overall cost measure. This can be least-squares difference from the sample mean (LS) or least absolute deviation from the sample median (LAD). A binary decision rule is found at each internal node by incrementally cycling through all the values of predictors and linear combinations of predictors in the learning data until a threshold value is found that gives the greatest reduction of error after the split. Each case branches to the left or right based on a threshold value of a rule computed with one or several predictors. Further node splitting continues until a terminal node is reached, after which no further reduction of error occurs. The final value assigned to a terminal node is the average or median value of the predictand for cases that fell into the node, depending on whether the LS

or LAD cost measure was used. In this study we used LS.

The procedure for finding an honest tree architecture consists of growing a large tree that perfectly fits all the response variable (predictand) data, then systematically eliminating nodes from the bottom upward (pruning) until the tree structure with the lowest error when tested with independent data is found. At each step of node elimination, an unbiased estimate of the error of the new tree is found by using it with a reserved independent dataset, or the error is estimated by cross validation (CV) for datasets of less than about three thousand cases. (Briefly, the CV procedure works as follows: A data sample is divided into a user-specified number of equal-sized bins each, insofar as possible, with about the same distribution as the entire sample. As each bin of reserved data is selected in turn, CART builds a decision tree to the current pruned complexity level with the remaining data not in the bin, drops the reserved data down the tree, and calculates the error of the fit to the data in the bin. The average error for the bins gives a "CV" error that is generally a good estimate of the error for truly independent data.) There is no well-defined minimum number of cases needed for CART, but as in any statistical data-fitting procedure, the more data available the more confidence one has in the generality of the result. It is possible, however, to obtain trees with as few as about 20 cases.

The "best" tree is considered to be one whose unbiased error is at or near the minimum. The most complex tree will generally have the highest unbiased error because the data have been overfit, that is, the data has been fit beyond the degree to which the predictors are really capable of fitting it (noise level). The error will usually slowly decrease as nodes are eliminated until a minimum is reached, then it will increase as nodes continue to be eliminated until only one is left. In this study, a tenfold (or 10 equal-sized bins) CV procedure was used to estimate the unbiased error of predictions by each tree, and the final trees selected had CV errors at the minimum or within one standard deviation of it. [This is the "1 SE rule" method for tree selection suggested by Breiman et al. (1984).]

Another measure of error in CART trees useful for comparison with results from other statistical methods is the "resubstitution error." This is simply the error of the fit of the CART tree to the learning data and is similar to, but not equivalent to, the "unexplained variance" measure used in linear regression. It gives an overly optimistic measure of the true error to be expected when using the tree with independent data, and the CV error estimate is a more realistic measure. For example, a large tree that perfectly fits all the data would have zero resubstitution error because even noise has been fit, but the CV error is likely very high because the tree would have difficulty fitting independent data. For easy comparison, the CV and resubstitution errors are usually expressed in tree-sequence summaries as a fraction of the initial variance in the learning data for LS or average deviation from the sample median for LAD.

CART tallies an ad hoc ranking of "predictor importance" on a scale of 0–100 based on the reduction of variance achieved when a predictor appears either in decision rules or as a surrogate predictor. The information here is very useful for separating out the predictors that are important from those that are not, although one should be liberal when interpreting relative ranking of predictors. The most "important" predictors generally appear early in the node-splitting process and so are related to the most direct segregation of high and low predictand values. These predictors will have considerable impact on the final outcome since an early branch down one side of the tree can lead to a much different answer from a branch down the other side. Other predictors deemed less important will often appear in the nodes farther down on one side or the other of the tree (i.e., after the early splits), and are related to the details present in the events that separate degrees of high or low values, thus they are still of considerable interest. Predictors with importance ranking less than about 5–10 are generally of little value, either because they are not important physically, or their signal is weak and is overcome by other predictors. However, caution should be used when interpreting low predictor importance rankings, particularly for predictors whose values are assigned by subjective interpretation, such as synoptic weather map type. A low importance may simply mean too much error is associated during the subjective assignment of the predictor value and the methodology for assigning the rule in each situation should be assessed. Also, rankings are given for the dataset as a whole, so low importance can be assigned to predictors that are important for splitting events in a few nodes far down the tree.

One of the advantages of CART is that correlation among predictors (parsimony) is not a problem as it is with other statistical methods. In fact, inclusion of correlated predictors sometimes gives a practical advantage in cases of missing data, since CART finds decision rules for surrogate predictors to be used when a required predictor is missing, and the surrogate predictor is often one correlated with the missing one.

When analyzing internal splitting rules, it is often useful to examine the "surrogate" and "competitor" rules found by CART. A surrogate rule is the rule that can be used to split an internal node when the usual rule cannot be calculated due to missing data. A competitor rule is the "next best" rule CART found for splitting a node. An hierarchy of these rules should be printed out with every node during analysis. They often give added insight into the role of predictors that may not appear in the tree structure because a better split has been found but are nevertheless important. Because surrogate and competitor rules are limited to a single predictor, they can be particularly useful for interpre-

tation when the best splitting rule is a linear combination of predictors.

REFERENCES

Atlas, L., R. Cole, Y. Muthusamy, A. Lippman, J. Connor, D. Park, M. El-Sharkawi, and R. J. Marks II, 1990: A performance comparison of trained multilayer perceptrons and trained classification trees. *Proc. IEEE,* **78,** 1614–1619.

Breiman, L., J. H. Freidman, R. A. Olshen, and C. J. Stone, 1984: *Classification and Regression Trees.* Wadsworth & Brooks/Cole, 358 pp.

Burrows, W. R., 1990: Tuned perfect prognosis forecasts of mesoscale snowfall for southern Ontario. *J. Geophys. Res.,* **95,** 2127–2141.

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

——, and R. A. Assel, 1992: Use of CART for diagnostic and prediction problems in the atmospheric sciences. Preprints, *12th Conf. on Probability and Statistics in the Atmospheric Sciences,* Toronto, ON, Canada, Amer. Meteor. Soc., 161–166.

Chung, Y. S., 1977: Ground-level ozone and regional transport of air pollutants. *J. Appl. Meteor.,* **16,** 1127–1136.

Heidorn, K., and D. Yap, 1986: A synoptic climatology for surface ozone concentrations in southern Ontario. *Atmos. Environ.,* **20,** 695–703.

Hudischewskyj, A. B., and T. E. Stockenius, 1991: Classification of Los Angeles basin ozone episodes on the basis of meteorological conditions. *Proc., 84th Annual Meeting Air & Waste Management Association,* Vancouver, BC, Canada, A&WMA, 23 pp.

Lelieveld, J., and P. J. Crutzen, 1990: Influences of cloud photochemical processes on tropospheric ozone. *Nature,* **343,** 227–233.

Lord, E. R., 1993: Forecasting daily maximum surface ozone concentrations in greater Vancouver and the Lower Fraser Valley. Atmospheric Environment Service, Pacific Region, Atmospheric Issues and Services Branch, Vancouver, BC, Canada, PAESS-93-3, 65 pp.

McKendry, I. G., 1993: Surface ozone in Montreal, Canada. *Atmos. Environ.* **27B,** 93–103.

MEP, 1990: A study of the peak ozone levels in the Toronto area. Final report submitted to the Ontario Ministry of the Environment. Prepared by the MEP Company, 40 pp. [ISBN 0-7729-9164-2.]

Mukammal, E. I., H. H. Neumann, and T. J. Gillespie, 1982: Meteorological conditions associated with ozone in southwestern Ontario, Canada. *Atmos. Environ.,* **16,** 2095–2106.

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

Ramage, C. S., 1993: Forecasting in meteorology. *Bull. Amer. Meteor. Soc.,* **74,** 1863–1871.

Robeson, S. M., and D. G. Steyn, 1990: Evaluation and comparison of statistical forecast models for daily maximum ozone concentrations, *Atmos. Environ.,* **24B,** 303–312.

Stockenius, T. E., 1991: A multivariate data analysis technique for assessing the influence of meteorological conditions on ozone concentration trends. *Proc., 84th Annual Meeting Air & Waste Management Association,* Vancouver, BC, Canada, A&WMA, 18 pp.

Yap, D., D. T. Ning, and W. Dong, 1988: An assessment of source contributions to the ozone concentrations in southern Ontario, 1979–1985. *Atmos. Environ.* **22,** 1161–1168.

Zeldin, M., and J. Cassmassi, 1978: Development of improved methods for predicting air quality levels in the South Coast Air Basin. Tech. Rep., Technology Service Corporation, Santa Monica, CA.