# The Use of Routine Weather Observations to Calculate Liquid Water Content in Summertime High-Elevation Fog

JOHN L. WALMSLEY,* WILLIAM R. BURROWS, AND ROBERT S. SCHEMENAUER*

*Atmospheric Environment Service, Downsview, Ontario, Canada*

(Manuscript received 3 November 1997, in final form 11 June 1998)

ABSTRACT

This paper represents a stage within a larger project to estimate acid ion deposition from cloud impacting on high-elevation forests. Acid ion deposition depends principally on three factors: the liquid water content (LWC), the ion concentration(s) in fog or cloud water, and the efficiency of the deposition process. In the present paper, the objective is to estimate LWC on Roundtop Mountain in southern Quebec from routine meteorological measurements at the Sherbrooke weather station.

After describing preliminary efforts, the methodology that was found to work best is presented. This scheme was a hybrid of applications of two statistical nonlinear regression schemes. First, the classification and regression trees (CART) algorithm was applied to predict the occurrence or nonoccurrence of fog at Roundtop. The algorithm produced by this application permitted the elimination of a large proportion of the data records for which fog was very unlikely to occur at Roundtop. The remaining data were then processed by a second application of CART to determine the predictors that are important for estimating LWC at Roundtop. Finally, these same remaining data were processed by the neuro-fuzzy inference systems (NFIS) algorithm to derive the final prediction algorithm. This hybrid method (CART–CART–NFIS) achieved a correlation coefficient of 0.810, with accuracies of 0.962 and 0.664 for the no-fog and fog events, respectively. (Corresponding threat scores were 0.916 and 0.530, respectively.) These measures of skill were significantly better than those obtained from initial estimates or from schemes that used CART alone.

Although optical cloud detector and LWC data are necessary for derivation of the fog-occurrence and LWC prediction algorithms, in the end those algorithms are applied to only the predictor data. Fog-occurrence and LWC data are not required, except for verification purposes. The algorithms and list of predictors still need to be tested to determine how widely applicable they are.

## 1. Introduction

Between 1985 and 1991, the Chemistry of High Elevation Fog (CHEF) experiment was conducted on three mountains in southern Quebec, Canada (Schemenauer et al. 1995). The CHEF project was linked with the Mountain Cloud Chemistry Project (MCCP) in the Appalachian Mountains of the United States. Measurements were made of liquid water content (LWC) as well as standard meteorological parameters (temperature, pressure, relative humidity, wind speed, wind direction, and precipitation). Fogwater and precipitation samples were collected and analyzed for the concentrations of various acid ions. The present paper is a continuation of a project, the earlier stages of which appeared in Bridgman et al. (1994), Walmsley et al. (1995, 1996a,b) and Urquizo et al. (1998).

Fog in its simplest definition is a cloud with its bas on the ground. Indeed, the fog events experienced o Roundtop Mountain are in the most part due to cloud passing over the mountain. From the reference point c the CHEF site on the mountain, however, the measure ments are most definitely made in fog.

Acid ion deposition depends principally on the LWC ion concentrations in fogwater, and efficiency of th deposition process. Previous studies at two high-ele vation [almost 1000 m above mean sea level (MSL) CHEF sites showed that the acid deposition from fo was at least as important as that from precipitation. I the present paper, the objective was to predict or est mate LWC in high-elevation fog. More specifically, th goal was to estimate LWC on Roundtop Mountain, Que bec, from routine meteorological measurements at th Sherbrooke weather station (see Fig. 1 and Table 1 The use of standard weather data is the first step i extending our predictive capability to areas where, an time periods when, fogwater data do not exist. In thi approach, we used two powerful statistical algorithms

* Retired.

*Corresponding author address:* Dr. John L. Walmsley, Atmospheric Environment Service, 4905 Dufferin Street, Downsview, ON M3H 5T4, Canada.
E-mail: john.walmsley@ec.gc.ca

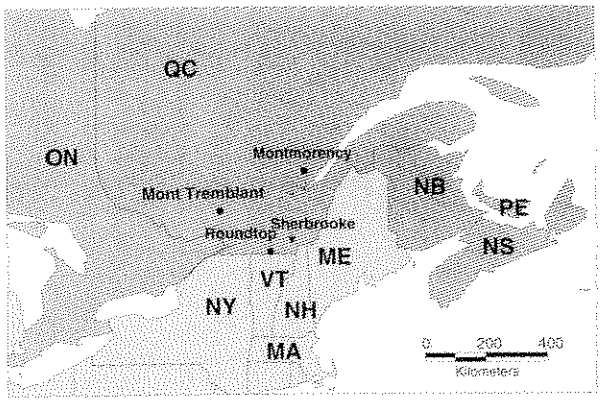


FIG. 1. Map of southeastern Canada and northeastern United States showing the three CHEF sites in southern Quebec and the Sherbrooke airport weather station, located 78 km northeast of the Roundtop Mountain CHEF site.

classification and regression trees (CART) (Breiman et al. 1984; Steinberg and Colla 1995) and neuro-fuzzy inference systems (NFIS) (Chiu 1994; Mathworks 1998). Both are methods of nonparametric, nonlinear regression, which are considerably different from traditional multiple linear regression schemes.

To apply LWC estimates from the method proposed in the present study to calculate fog deposition in complex terrain, it will be necessary to predict the frequency of fog occurrence at a given altitude with an absolute accuracy of 10% or better. This is because fog frequencies typically fall in the range of 0%–50% for mountain forests. Any higher uncertainty in the fog frequency would produce little improvement over simply calculating the wet deposition from the precipitation component alone. Similarly, the calculations will have to be done with altitude steps of 100 m to account for changing forest areas with altitude.

In the remaining part of this introductory section, we describe the CART and NFIS algorithms, as well as a hybrid scheme that makes use of both algorithms. Sections 2 and 3 describe the acquisition and preparation of cloud detector and LWC data used in this study. Sections 4 and 5 outline preliminary attempts to estimate LWC; section 6 presents the final scheme. We summarize our results and present conclusions in section 7.

### a. CART

The CART program is ideally suited to the problem of analyzing a large amount of data consisting of an observed value that one wishes to be able to estimate or predict (i.e., the *predictand*) and a number of possib *predictors*. The predictand and predictors may take continuous values or may be classified into a finite nu ber of categories. Output is at discrete values. A CA fit of the data has the form of a decision tree and piecewise continuous. It uses a procedure of minimizi variance at each "splitting" of the data at internal nod of the tree. During this process, CART selects the r evant predictors from what may be a large number potential ones. A more detailed description of CA appears in a recent application (Burrows 1997).

An important CART input option is the method error estimation. In the present study, we selected t test set method. Initially, one year was chosen as a lea ing set and another as the test set. Later, four years data were combined in one large dataset and a specifi percentage of the total dataset was reserved for testi purposes.

The following is a conceptual description of h CART operates, from the viewpoint of the user. CA processes the data and attempts to find a tree, consisti of a number of nodes, that minimizes the predicti error. Node 1 of the tree is the entire dataset. CA builds a tree by splitting the data, beginning with no 1, into two new nodes. These nodes may also be sp a node that is not split is referred to as a terminal no (TN). Each split is accomplished according to whet a predictor or a linear combination of predictors is abo or below a threshold value, examples of which shown in section 5. For simplicity in this paper, lin combinations of predictors were not permitted, so splitting was done using a single split variable. In ea application, we show the tree, a table giving details the splitting and nodes, and a table giving the termi nodes and their corresponding predicted values.

As we had many small observed values of the p dictand, we initially tested the logarithmic transform tion option in CART to produce a more even distributi of values. We then abandoned that approach and ha now replaced it with a method, to be described belo to eliminate most of the events in which fog was predicted to occur at the Roundtop observation site.

### b. NFIS

Unlike CART, NFIS does not eliminate predicto however, it does provide a continuous-response form output. NFIS uses subtractive clustering to form clus centers. Formation of cluster centers is controlled t certain extent by the user's choice of four input para

TABLE 1. Locations and elevations of the CHEF ridge site on Roundtop Mountain and the Sherbrooke airport weather station.

| Site | Elevation (m MSL) | Lat (N) | Long (W) | UTM east (m) | UTM north (m) |
|---|---|---|---|---|---|
| Roundtop Ridge | 845 | 45°05′17″ | 72°33′09″ | 692 625 | 4 995 400 |
| Sherbrooke A | 238 | 45°26′ | 71°41′ | 790 107 | 5 034 375 |

eters. The *radius of influence* controls the size (in multidimensional vector space) of each cluster. The *squash* factor controls the degree of overlap of clusters. The remaining two parameters control admission and rejection of new cluster centers. Fuzzy membership of data points in the cluster centers is assigned by Gaussian membership functions. Membership is "fuzzy" in the sense that a data point may have membership in more than one cluster. Each cluster center is regarded as a basis for fuzzy rules, which are formulated as linear equations in the predictors. Training data are used to determine the coefficients of these equations, resulting in a linear least squares estimation problem that takes the form of a matrix equation: $\mathbf{AX} = \mathbf{B}$. A recursive procedure is used to solve for $\mathbf{X}$, which is the solution vector containing the coefficients of all the linear equations that represent the fuzzy rules.

### c. CART–NFIS hybrid method

In practice, we used CART as a preprocessing step to select the predictors for use by NFIS. Application of NFIS then provided continuous output (unlike CART's discretely spaced values) based on predictors that CART had found to be significant. In this manner, we combined the strengths of the two methods. We found that the hybrid method, to be described in more detail in section 6, gave significantly better results than either CART or NFIS alone.

## 2. Measurement program

A fog-measuring device, FMD-04R (Advance Electronic Sub-Systems, Inc., Kettleby, Ontario, Canada), was used to measure LWC. The device, hereafter the FMD, was based on a design described by Gerber (1984). It consisted of a planar circular photodiode placed perpendicular to and coaxial with a narrow collimated beam of light. The beam itself was captured by a light trap. The FMD sensed the light scattered out of the beam direction through a small angular range. In principle, the scattered flux $F$ (W) could be related to the beam radiant flux $I$ (W) and the liquid water content $W$ (g m$^{-3}$) by Gerber's relation:

$$F = 0.008\ 25IW. \tag{1}$$

In practice, however, the FMD output $\mathcal{F}$ (mV) was first corrected for zero-drift $c$ (mV) and was then used to derive $W$ using a calibration curve (M. Wasey 1997, personal communication)

$$W = m(\mathcal{F} - c) + b, \tag{2}$$

where $m = 0.021$ g m$^{-3}$ mV$^{-1}$, $b = 0.0029$ g m$^{-3}$, and $c \sim 50$–$100$ mV. The value of $c$ was checked and recorded daily (in clear-air conditions) by the operator. In the case of small changes, the value was not reset, but the data were later corrected by subtracting the zero value. In the case of substantial changes in zero value, the FMD was shut off, the optical path and opti cleaned, the electronics checked, and the instrument r started. The zero value was then noted before new da were collected. This daily protocol, in addition to examination of the time-series plot for the month, e abled small changes from the zero value to be taken valid measurements of LWC. As part of the processi of the raw data, any remaining zero-drift correction w done somewhat subjectively from time-series plots $\mathcal{F}$. The period over which a given zero correction w applied depended on the judgment of the analyst. Ther fore several values of $c$ could be used between mai tenance and adjustment visits to the measurement sit

The values of $m$ and $b$ in (2) were determined fro a wind-tunnel comparison with a PMS forward scatte ing spectrometer probe (FSSP). Equation (2) is intend for use with drops of size 6–10 $\mu$m and is valid for mV $< \mathcal{F} - c <$ 40 mV (i.e., 0.02 g m$^{-3}$ $< W <$ 0 g m$^{-3}$, approximately). Hourly averaged LWC data o tained from this device were available for vario lengths of time between April and October in the perio 1988–91.

In addition to the FMD, hourly averaged data we available for the years 1989–91 from an optical clou detector (OCD) designed to indicate the presence absence of fog. The OCD (model MCD-05; now know as the Fog Sentinel, Series DH, Advanced Sensor Tec nologies, Inc., Abingdon, MD) was based on a prototy developed at the Netherlands Energy Research Fou dation. An infrared signal is transmitted toward a ph toelectric receiver. In the absence of cloud or fog, t signal is blocked by a "blocker post" and is not detecte at the receiver. When cloud or fog is present, the infrar beam is near-forward scattered around the blocker po enabling the receiver to detect the signal. At Roundto the OCD was calibrated to detect fog when the LW exceeded approximately 0.07 g m$^{-3}$ (M. Wasey 199 personal communication).

The size distribution of the fog droplets can affe the response of the FMD and the OCD at values ne the threshold responses of the instruments. In additio the absolute value of the LWC measured by the FM and reflected in the "turn on" setting of the OCD only as good as the calibration of the PMS FSSP again which the FMD was compared. At the CHEF site, hov ever, the OCD signaled the presence of fog when t FMD indicated LWC values of about 0.07 g m$^{-3}$. The are times when the agreement between the instrumen is poorer than at other times. Since each hourly avera is made up of many individual measurements, howev the instruments should be compatible when considerin hourly values.

When the CHEF site was below cloud base, and th not in fog, and precipitation was falling, neither t FMD nor the OCD had a significant response to t precipitation. When the cloud base was below the sar pling elevation, and the site was in fog, both instrumen responded to the fog.

TABLE 2. LWC and FD hourly data availability and processing.

| Year | No. records[a] | LWC available[b] | LWC ≤0.6[c] | FD available[d] | FD <200[e] | FD Mid[f] | FD ≥2800[g] | Percent FD mid.[h] | Processing[i] | Processing percent[j] |
|---|---|---|---|---|---|---|---|---|---|---|
| 1988 | 2894 | 2887 | 2886 | 0 | 0 | 0 | 0 | 0.0 | 2886 | 100.0 |
| 1989 | 3079 | 2144 | 2144 | 1699 | 1269 | 123 | 110 | 7.2 | 2377 | 110.9 |
| 1990 | 807 | 755 | 755 | 782 | 730 | 123 | 92 | 15.7 | 683 | 90.5 |
| 1991 | 3579 | 1542 | 1532 | 1612 | 1538 | 184 | 209 | 11.4 | 1411 | 91.5 |
| 1988–91 | 10 359 | 7328 | 7317 | 4093 | 3252 | 430 | 411 | 10.5 | 7357 | 100.4 |

[a] No. records — Total observational period (h)
[b] LWC available — Number of LWC hourly records
[c] LWC ≤ 0.6 — Number of valid LWC records
[d] FD available — Number of FD hourly records
[e] FD < 200 — $N_{FD} < 200$
[f] FD mid. — $200 \leq N_{FD} < 2800$
[g] FD ≥ 2800 — $N_{FD} \geq 2800$
[h] Percent FD mid. — FD mid./FD available
[i] Processing — Number LWC records after FD processing
[j] Processing percent — Processing/LWC available

Observations began
1100 LST 17 May 1988
1000 LST 11 May 1989
1300 LST 3 Oct 1990
1100 LST 24 May 1991

Data from the Sherbrooke weather station used in this study are available in the Canadian Weather Energy and Engineering Data Sets (CWEEDS) compiled by the Atmospheric Environment Service in 1993. A number of the CART predictors used in the estimation of LWC were obtained directly from the CWEEDS files. These included solar radiation variables, cloud ceiling height, sky condition, visibility, weather elements, station pressure, air temperature, dewpoint, wind speed, and wind direction. Other predictors were derived directly from the CWEEDS data and from date, time, location, and elevation. These included dewpoint depression, relative humidity, air density, lifting condensation level (LCL), solar declination, solar altitude and azimuth angles, and a day/night category variable. A preliminary estimate of LWC at the Roundtop observation site (845 m MSL) was made by integrating upward from the LCL assuming, as in Walmsley et al. (1996a), that 38% of the condensed liquid water was retained during the ascent. This estimate was then adjusted based on a "probability of cloud" as determined from CWEEDS weather elements, sky condition, cloud ceiling height, and an estimate of thermal stability.

## 3. Data preparation

For convenience, the output from the OCD will be referred to as fog detector (FD) data. An examination of the hourly averaged values of FD revealed that in only about 10% of the cases did the OCD detect fog for part of the hour. The rest of the time, it either detected fog for the whole hour (FD ≥ 2800) or (in the majority of the cases) did not detect it at all (FD < 200). To simplify the problem, it was decided to eliminate from further consideration the hourly LWC records with a mixture of fog and no fog (200 ≤ FD < 2800). At the same time, when FD < 200, the LWC was set to zero, regardless of its recorded value or whether it was originally missing. When FD was missing, the LWC was left unchanged. Table 2 summarizes the LWC ar FD data availability and the screening process. It shou be noted that in 1989, the number of LWC records aft processing with the FD data was increased due to tl resetting of some missing LWC values to zero. In 198 one observed value of LWC exceeded 0.6 g $m^{-3}$; th was considered doubtful and was removed. There we 10 similar cases in 1991. The FD screening proce yielded a net increase of 40 LWC records (i.e., fro 7317 to 7357). The FD data were available for 3567 the 7357 records to be processed by CART and NFI Also shown in Table 2 is the date and time of the sta of operations and the total number of hours of operatic for each of the four years.

An estimate of fog frequency can be made from Tab 2, which indicates that 4093 hourly records of FD we available. (This number is larger than the 3567 prev ously mentioned, as the latter included only those hou for which LWC was also available.) Of the 4093 hour only 411 (or 10%) had fog for the complete hour; hov ever, there were an additional 430 hours (or 11%) which fog occurred for part of the hour. This gives upper estimate of fog occurrence at 21%; however, more reasonable approach would be to take half of tl part-hour data (or 5.5%). This would give a more r alistic estimate of fog occurrence at about 15% as average over the time period used in this study.

## 4. Preliminary LWC estimates

### *a. Calculation of LWC from Sherbrooke data*

As a first step, the LCL was calculated from the ter perature $T$ and dewpoint $T_d$ at Sherbrooke. Next, v used the procedure described in Walmsley et al. (1996a assuming retention of 38% of the liquid condensate, integrate upward to the height of the Roundtop CHI site, where a first estimate of the LWC was determine In principle, this is a fairly straightforward applicati

of several thermodynamic relations. In practice, however, the calculation scheme is somewhat complicated by the implicit nature of those relations. Figure 2a shows a comparison between the predicted and observed values for the 1988 season. As can be seen, there is a tendency to overestimate the LWC. Furthermore, there is a great deal of scatter, particularly when the observed LWC is near zero. There is a poor Pearson linear correlation coefficient of 0.321.

### *b. Adjusted LWC prediction*

Figure 2b shows the results of an attempt to improve the LWC estimate using other meteorological data at Sherbrooke: sky condition, ceiling height, weather elements, and atmospheric stability. The stability class used was calculated from the STAR program. The tendency to overpredict has essentially been eliminated and the correlation coefficient has been improved to 0.479, but there is still a large amount of scatter.

### *c. Smoothing of data*

It was thought that the scatter in Fig. 2b may have been due to the use of hourly values at Sherbrooke to predict hourly LWC at Roundtop, 78 km away, that is, there may be a lag between the time of the weather station observations and the time of the corresponding LWC measurements. Using a three-point smoother on both the observed and predicted LWC would possibly reduce such lag effects. Figure 2c shows the results of this exercise. The correlation coefficient is increased to 0.531, and the scatter is reduced slightly, but it is still unacceptable.

## 5. Preliminary CART–NFIS applications

### *a. 1988 dataset*

Initial CART runs on the 1988 dataset did not show remarkable improvement over the results of Figs. 2b,c. Difficulties became even more apparent when additional years of data (1989–91) became available. Attempts were made to generate a CART tree from each of the four years of data in turn and to use the resulting algorithm to predict for the remaining three years. None of the 12 such applications showed much promise.

It was noticed that most of the CART trees produced during the above-mentioned tests used the cloud ceiling height at Sherbrooke as the predictor at the first splitting of the data. Accordingly, it was decided to perform further tests with the 1988 data. The dataset was split prior to processing by CART according to whether or not the observed ceiling height converted to height above MSL (CMSL) was less than 910 m. (A threshold of 910 mb was selected by CART in early tests on the 1988 data.) The two data subsets that resulted were then processed by CART to determine appropriate predictors in each case. Those predictors were then used by NFIS to p duce the output shown in Fig. 3.

In Fig. 3a, the results for the 286 hourly calculati when CMSL $\leq$ 910 m are displayed. The linear c relation coefficient is a respectable 0.875. The best linear regression line lies close to the perfect-fit (1 line. Its slope and y intercept are heavily influenced the congregation of points near the y axis where LWC was calculated to be about 0.05 g $m^{-3}$. Four the highest LWC values straddle the perfect-fit line, s gesting that the fit would be remarkably good were not for a tendency to predict about 0.05 g $m^{-3}$ wh the observed values were less than 0.02 g $m^{-3}$. In su mary, Fig. 3a shows that by eliminating the cases high ceilings, most of the no-fog events were effectiv eliminated. The remaining events, which contain a m ture of fog and no-fog cases, are seen to be well p dicted.

Figure 3b shows the results when high ceilir (CMSL $>$ 910 m), including unlimited ceiling (i.e., sc tered cloud or clear skies) were observed at Sherbroo There were 2600 such hours when LWC data were a available. These cases contain a large number of no-f events (clustered near the y axis) and a relatively sm number of fog events. The latter events are mostly u derpredicted, resulting in an overall poor correlation c efficient of 0.474. The scatter about the best-fit li however, is not large except near the y axis where large number of points are congregated.

Figure 3c combines the points from Figs. 3a,b. T correlation coefficient is 0.782, which is considered a ceptable; however, the scatter is still a concern. Ne ertheless, it would seem that prescreening of the d can lead to better predictability of the LWC. This ic will be pursued below.

### *b. 1988–91 dataset*

First, however, an initial investigation will be ma on a dataset constructed from the combined 1988, 19 1990, and 1991 data without prescreening. Figure 4 p sents the tree produced by CART using 80% of t 1988–91 data records as a learning set. The dataset w split four times, producing five terminal nodes. A c scription of CART's tree-building process appears section 1a; details of the splitting are given in Table Terminal-node information is in Table 4 and is discuss below. Predictor DTDZ was obtained from the observ temperature at Sherbrooke and the estimated tempe ture at the Roundtop CHEF site calculated from t LCL–LWC procedure (section 4a). Predictor JDSUM derived from the month and day. The remaining tv predictors are from meteorological observations at Sh brooke. The ceiling height predictor could possibly made more universal by subtracting the elevation of t Roundtop CHEF site (845 m).

Three of the predictors selected by CART are phy ically meaningful. Intuitively, the observed ceili

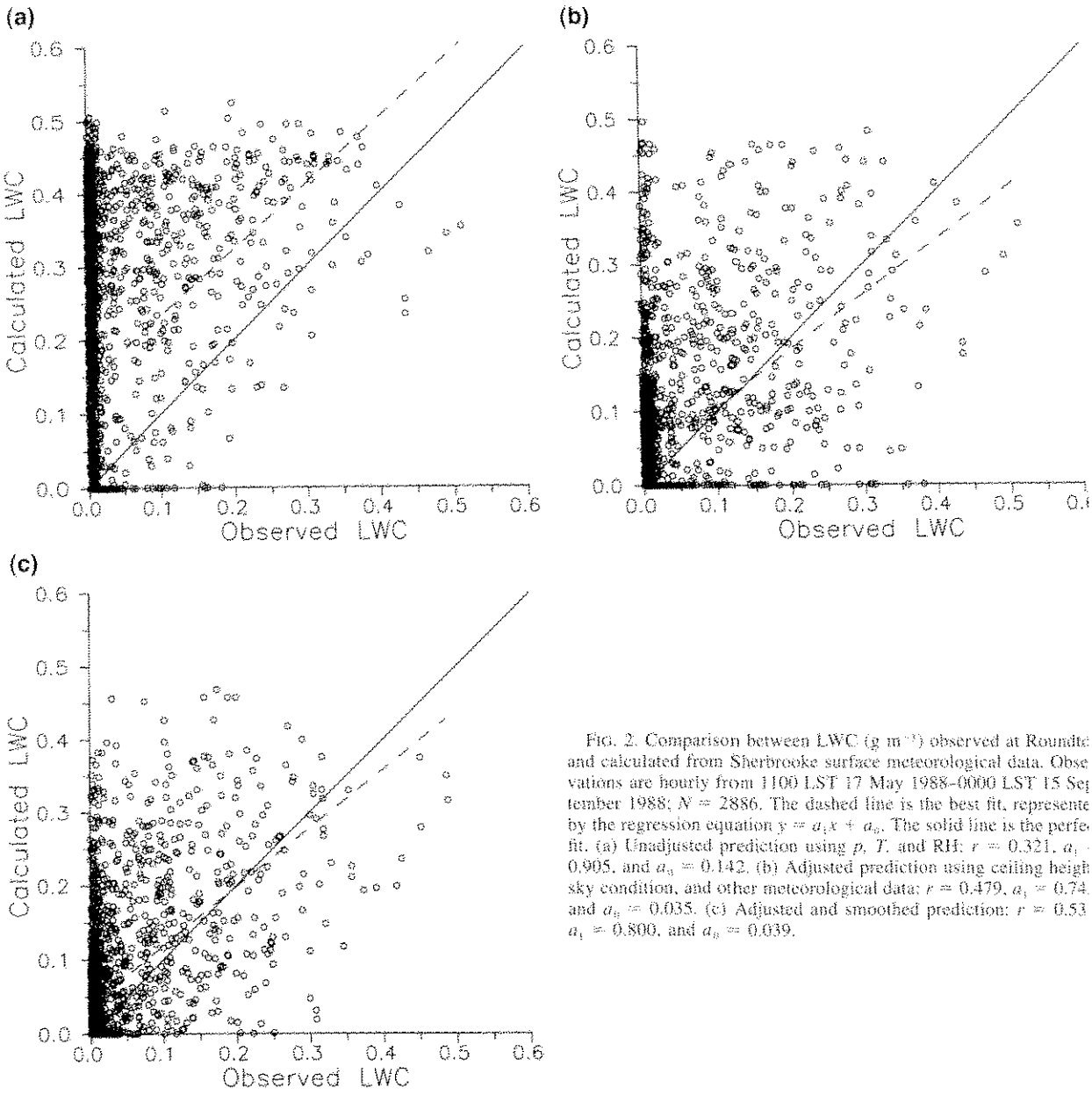


FIG. 2. Comparison between LWC (g m$^{-3}$) observed at Roundto and calculated from Sherbrooke surface meteorological data. Obse vations are hourly from 1100 LST 17 May 1988–0000 LST 15 Se tember 1988; $N = 2886$. The dashed line is the best fit, represente by the regression equation $y = a_1x + a_0$. The solid line is the perfe fit. (a) Unadjusted prediction using $p$, $T$, and RH: $r = 0.321$, $a_1$ 0.905, and $a_0 = 0.142$. (b) Adjusted prediction using ceiling heigh sky condition, and other meteorological data: $r = 0.479$, $a_1 = 0.74$ and $a_0 = 0.035$. (c) Adjusted and smoothed prediction: $r = 0.53$ $a_1 = 0.800$, and $a_0 = 0.039$.

height should be a good indicator of the presence or absence of fog at the Roundtop CHEF site and, if present, the amount of LWC in the fog. The vertical temperature gradient is an indicator of atmospheric stability, which along with atmospheric pressure, is often a good indicator of weather conditions. The use of Julian day is more obscure; nevertheless, it did cause 553 cases to be split into subsets of 82 and 471 cases, respectively, as shown in Tables 3 and 4. Thus, JDSUM was not simply used to isolate a few outlying points. One obvious conclusion resulting from an examination of Fig. 4 and Table 3 is that human intervention is very unlikely to have arrived at CART's choice of predictors, let alone its threshold values for those predictors.

Table 4 lists the average values and standard devi tions at each terminal node, both for the dataset and t test set (i.e., the randomly selected 20% of the 1988 91 data that was not used to derive the tree but w reserved for testing purposes). The average values i the learning set, therefore, are the predictions again which the corresponding observed test set values shou be compared. This comparison is also shown in Fig. . where it can be seen that the prediction of the averag value at each node is very good; however, there is large standard deviation in the test set data at all nod except the lowest valued (Node 5 in Table 4).

Figure 6 displays this information in a different fas ion: the plotting of all 7357 points from the 1988–9

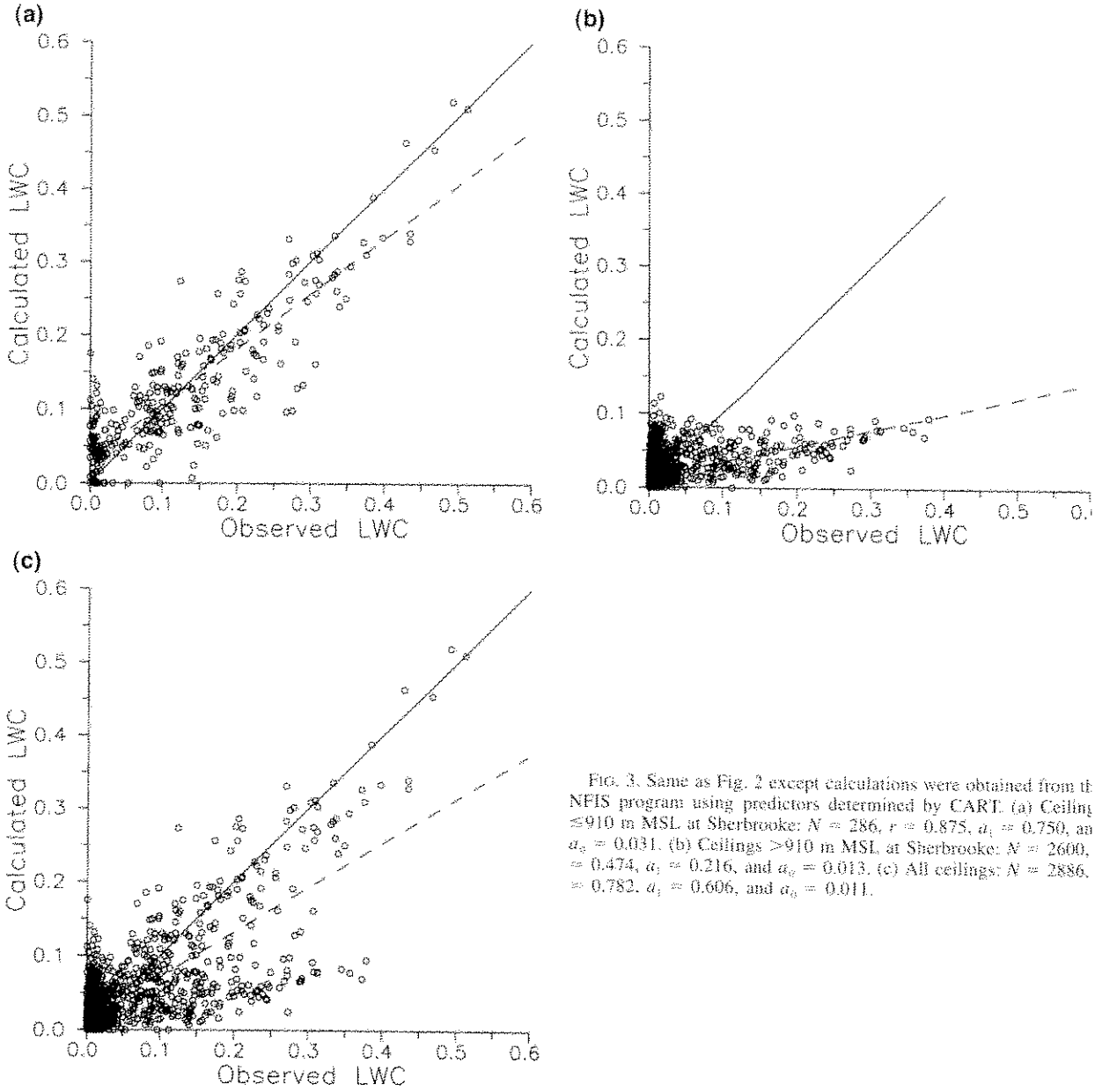


FIG. 3. Same as Fig. 2 except calculations were obtained from th NFIS program using predictors determined by CART. (a) Ceiling ≤910 m MSL at Sherbrooke: $N = 286$, $r = 0.875$, $a_1 = 0.750$, an $a_0 = 0.031$. (b) Ceilings >910 m MSL at Sherbrooke: $N = 2600$, = 0.474, $a_1 = 0.216$, and $a_0 = 0.013$. (c) All ceilings: $N = 2886$, = 0.782, $a_1 = 0.606$, and $a_0 = 0.011$.

dataset. Unlike the NFIS plots of Fig. 3, the results produced by CART are only available at the five terminal nodes, hence the five horizontal lines of data points. The scatter that has persisted throughout this project is again evident here.

Table 5 presents another representation of the CART results for LWC prediction during 1988–91. Both the observed and predicted values have been separated into two bins at 0.04 g m$^{-3}$, an estimate of the LWC threshold above which fog is visible. The binned data are presented in the form of a 2 × 2 contingency table. The bottom row of totals shows that low values of LWC were observed in 6477 cases, or 88% of the time, whereas CART predicted low values in 6184 cases, or 84% of the time. (These 6184 cases represent the sum of th cases in the learning and test sets in Node 5, Table 4. More significant, however, is that the prediction of lov LWC had an accuracy rate of 0.952 (i.e., 5888 case out of 6184). Here, the prediction accuracy or post agreement (Stanski et al. 1990) is defined as

$$A_i = n_{ij}/r_i, \qquad i = j, \tag{3}$$

where $i$ is the row number (predictions), $j$ is the colum number (observations), $n_{ij}$ is the number of cases in cel $(i, j)$, and $r_i$ is the total of the $i$th row.

Predicting the infrequent event, on the other hand proved more difficult. The prediction accuracy for LW( > 0.04 g m$^{-3}$ was only 589 cases out of 1173, or 0.498

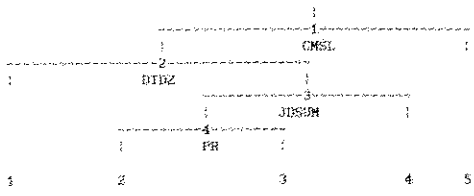


FIG. 4. CART regression tree diagram for LWC prediction derived from a learning set ($N$ = 5847) representing 80% of the 1988–91 data. The split variable is indicated at each of the four internal nodes (see Table 3). Terminal node information appears in Table 4.

TABLE 4. Terminal node information for LWC (g m$^{-3}$) predictic by CART from the 1988–91 data.

| | Learning set | | | Test set | | |
|---|---|---|---|---|---|---|
| Node | Cases | Avg | Std dev | Cases | Avg | Std de |
| 1 | 384 | 0.049 | 0.085 | 100 | 0.076 | 0.12( |
| 2 | 27 | 0.424 | 0.140 | 5 | 0.409 | 0.18: |
| 3 | 55 | 0.177 | 0.183 | 10 | 0.176 | 0.17( |
| 4 | 471 | 0.129 | 0.125 | 121 | 0.115 | 0.11: |
| 5 | 4910 | 0.011 | 0.041 | 1274 | 0.009 | 0.03: |
| Total | 5847 | | | 1510 | | |

These results are typical for predictions of rare events (Matthews 1997), demonstrating that an extremely accurate method is needed to achieve respectable scores.

Also shown in Table 5 is the threat score or critical success index (Stanski et al. 1990), defined as

$$T_i = n_{ij}/(r_i + c_j - n_{ij}), \qquad i = j, \tag{4}$$

where $c_j$ is the total of the $j$th column. It can be seen from (4) that $T_i$ ranges from 0 (none of the predictions of the $i$th category were correct) to 1 (all of the predictions of the $i$th category were correct). This statistic is sensitive to both missed events and false alarms. It is more representative of accuracy when rare events are involved than either the probability of detection

$$\mathrm{POD}_i = n_{ij}/c_j, \qquad i = j, \tag{5}$$

or the false alarm rate, $\mathrm{FAR}_i = 1 - A_i$ (Stanski et al. 1990).

The total accuracy or percent correct (Stanski et al. 1990) appears in the bottom-right box of each of the three contingency tables. It is computed as the sum of the diagonal elements divided by the total number of cases, $N$:

$$\text{Percent correct} = n_{ij}/N, \tag{6}$$

where the summation convention is in effect: for example, (5888 + 584)/7357.

It should be noted that the threshold of 0.04 g m$^{-3}$ in Table 5 was originally derived somewhat subjectively from inspection of Figs. 2 and 3. It is apparent that a large percentage of observed values are clustered below this value. (Table 5 confirms that 88% of the data lies at or below 0.04 g m$^{-3}$.) The value of 0.04 g m$^{-3}$ a threshold for fog events is also related to the visibil threshold specified in both the CHEF and the MC studies. In these major studies, the specialized fog c lectors were exposed when the visibility dropped bel 1 km. This is a very light fog event and corresponds measured LWC in the range of 0.04–0.05 g m$^{-3}$. Bel these LWC values, one sees little or no wetting of ve etation, and the collection of fogwater samples by t Teflon stringed collectors is nearly impossible. The fore, the instrumentation was designed to function w with a dynamic range starting at about 0.05 g m$^{-3}$, the measurements of most importance fall in the ran of 0.05 to about 0.5 g m$^{-3}$.

The results of Table 5 suggest that if an accur method can be derived to predict low LWC, then ma cases could be eliminated from the dataset. The reduc dataset would contain a much larger percentage of LV

TABLE 3. Learning set node information for LWC prediction by CART from the 1988–91 data. Terminal nodes are shown in bold figures. The tree structure is shown in Fig. 4.

| Node | Cases | Split variable | Threshold | Units | Left node | Right node |
|---|---|---|---|---|---|---|
| 1 | 5847 | CMSL[a] | 1123 | m | 2 | **5** |
| 2 | 937 | DTDZ[b] | −7.250 | K km$^{-1}$ | **1** | 3 |
| 3 | 553 | JDSUM[c] | −18 | day | 4 | **4** |
| 4 | 82 | PR[d] | 98.015 | kPa | **2** | **3** |

[a] CMSL, Cloud ceiling height above MSL
[b] DTDZ, Vertical gradient of temperature
[c] JDSUM, Julian day relative to summer solstice
[d] PR, Station pressure

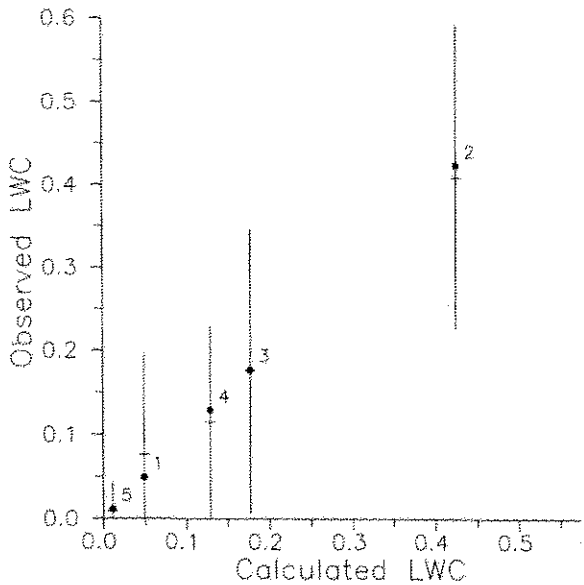


FIG. 5. CART prediction of LWC (g m$^{-3}$) at Roundtop in 19 91 for the test set, $N$ = 1510. Each terminal node is shown wit corresponding predicted value from the learning set, $N$ = 5847 (T 4). (Connecting the closed circles ● gives the 1:1 perfect-fit li Crosses (+) and vertical bars show the mean and std dev from test set (see Table 4). Note that the abscissa and ordinate axes reversed from their orientations in Figs. 2 and 3.

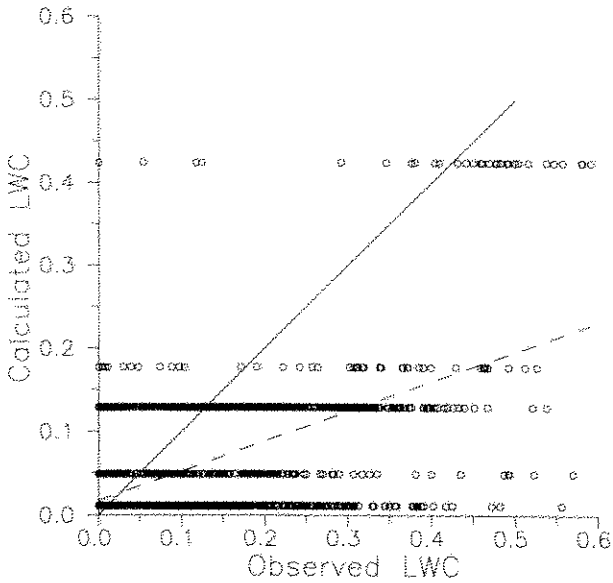


FIG. 6. CART prediction of LWC (g m$^{-3}$) at Roundtop for the complete 1988–91 dataset, $N = 7357$. Each predicted value is at one of the five terminal node values from the learning set (see Table 4). The dashed line is the best fit, represented by the regression equation $y = a_1x + a_0$. The solid line is the perfect fit: $r = 0.596$, $a_1 = 0.357$, and $a_0 = 0.017$ g m$^{-3}$.

$\geq 0.04$ g m$^{-3}$ than the complete dataset, thus giving a possibility for improved prediction accuracy. This idea will be pursued in the following section.

## 6. Final CART–NFIS applications

### *a. Optical cloud detector (OCD) data*

The original FD data were converted to FD categories or classes according to the following criteria: class 0 when FD $< 200$; class 1 when FD $\geq 2800$; otherwise class 9 (i.e., "missing"). As mentioned in section 3, out of 7357 original records (Table 2), 3567 were classified as 0 or 1. These records were processed by CART using the classification option to predict the absence (class 0) or presence (class 1) of fog at the Roundtop CHEF site.

It was subsequent to the choice of the threshold of 0.04 g m$^{-3}$ in Table 5 that it was decided to use the OCD data to identify hours during which fog occurred

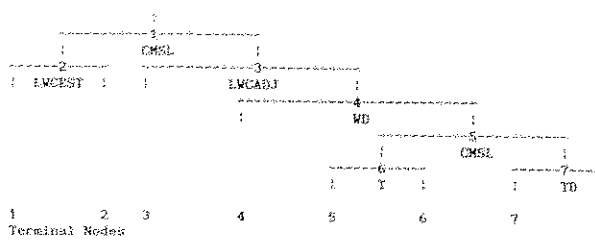


FIG. 7. CART regression tree diagram for prediction of fog o currence at the Roundtop CHEF site. The split variable is indicat at each of the seven internal nodes (see Table 6). Terminal no information appears in Table 7.

0% and 100% of the time. The estimated OCD detectio limit of 0.07 g m$^{-3}$ (see section 2) therefore correspon to approximately FD = 2800. A value of 0.04 g m probably corresponds to about FD = 200, which is t threshold indicating no fog throughout the hour. T range between 0.04 and 0.07 g m$^{-3}$ (FD from 200 2800) probably contains cases in which fog occurre for part of the hour. As shown in section 3, only 11 of the observations fell in this range during the stud period.

### *b. CART prediction of fog occurrence*

Figure 7 shows the tree from this CART run (hereaft referred to as run 1). CART's tree-building process described in section 1a; details of the splitting are give in Table 6. Terminal-node information is in Table 7 an is discussed below. Four of the six predictors used a meteorological variables observed at Sherbrooke. Clou ceiling height converted to MSL height (CMSL) w used at nodes 1 and 5. Wind direction (WD), tempe ature ($T$) and dewpoint (TD) were the other meteor logical predictors. Predictor LWCEST is the first est mate of LWC made for Roundtop (section 4a; Fig. 2a whereas LWDADJ is the refined estimate (section 4 Fig. 2b).

In Table 7, terminal node results are presented fro CART run 1. Because this was a classification, rath than a regression run, the format is different from Tab 4. A proportion of the data (20%) was reserved for te purposes. Of the eight terminal nodes, five were cla sified as 0 (no fog) and the remaining three as 1 (fog

TABLE 5. Contingency table for CART prediction of LWC (g m$^{-3}$) at Roundtop. The learning set was a randomly selected 79.5% ($N = 5847$) of the 1988–91 dataset ($N = 7357$). The CART-derived algorithm was then applied to the entire dataset. The linear correlation coefficient is $r = 0.596$.

| Predicted LWC | Observed LWC 0–0.04 | Observed LWC >0.04 | Total | Accuracy | Threat |
|---|---|---|---|---|---|
| 0–0.04 | 5888 | 296 | 6184 | 0.952 | 0.869 |
| >0.04 | 589 | 584 | 1173 | 0.498 | 0.398 |
| Total | 6477 | 880 | 7357 | 0.880 | |

TABLE 6. Node information for prediction of fog occurrence. Te minal nodes are shown in bold figures. The tree structure is show in Fig. 7.

| Node | Cases | Split variable | Threshold | Units | Left node | Right node |
|---|---|---|---|---|---|---|
| 1 | 2854 | CMSL | 1408 | m | 2 | 3 |
| 2 | 647 | LWCEST | 0.023 | g m$^{-3}$ | **1** | **2** |
| 3 | 2207 | LWCADJ | 0.004 | g m$^{-3}$ | **3** | 4 |
| 4 | 791 | WD | 185 | deg | 4 | 5 |
| 5 | 477 | CMSL | 4138 | m | 6 | 7 |
| 6 | 132 | T | 13.1 | °C | **5** | **6** |
| 7 | 345 | TD | −1.9 | °C | **7** | **8** |

TABLE 7. Terminal node information for prediction of fog occurrence (class 0 = no fog; class 1 = fog).

| | | Learning set | | | Test set | | |
|---|---|---|---|---|---|---|---|
| Node | Class | Cases | 0 | 1 | Cases | 0 | 1 |
| 1 | 0 | 183 | 179 | 4 | 36 | 35 | 1 |
| 2 | 1 | 464 | 254 | 210 | 115 | 61 | 54 |
| 3 | 0 | 1416 | 1411 | 5 | 364 | 363 | 1 |
| 4 | 0 | 314 | 311 | 3 | 85 | 85 | 0 |
| 5 | 1 | 62 | 44 | 18 | 13 | 10 | 3 |
| 6 | 0 | 70 | 67 | 3 | 22 | 22 | 0 |
| 7 | 1 | 29 | 23 | 6 | 6 | 4 | 2 |
| 8 | 0 | 316 | 312 | 4 | 72 | 71 | 1 |
| Total | | 2854 | 2601 | 253 | 713 | 651 | 62 |
| Percent | | 100 | 91 | 9 | 100 | 91 | 9 |

The great majority (91%) of the cases in both the learning and test sets were observed to fall in class 0. Each terminal node may be examined individually for accuracy of the prediction. At terminal node 1, for example, 179 cases out of 183 in the learning set, or 98%, were predicted correctly to fall in class 0. The corresponding results for the test set were 35 out of 36, or 97%. At terminal node 2, however, only 210 cases out of 464 in the learning set, or 45%, were predicted correctly to fall in class 1. The corresponding accuracy in the test set was 54 out of 115, or 47%.

As can be seen, the observed cloud ceiling height is a strong predictor. Unless the total cloud amount exceeds 0.5, the observed ceiling is reported as "unlimited" and coded as a large number. When a ceiling is present, the value coded gives the height of the cloud base of the layer at which the accumulated cloud amounts (from the lowest layer upward) first exceeds 0.5. Hence, the lower the value, the more likely that the cloud cover at Roundtop exceeds 0.5 at and below the reported ceiling height and the more likely that fog would have been observed at the Roundtop CHEF site. An examination of Fig. 7 and Table 6 shows that a value of CMSL $<$ 1408 m leads to terminal nodes 1 or 2 (i.e., class 0 or 1, respectively), depending on the value of LWCEST; whereas CMSL $\geq$ 1408 m leads to the other six terminal nodes (four of which are class 0).

Our first two guesses of the LWC at Roundtop are useful predictors of whether or not fog occurred at the CHEF site. Figure 7 and Table 6 show that LWCEST is used to split internal node 2 into terminal nodes 1 and 2 (classes 0 and 1, respectively) at the threshold value of 0.023 g m$^{-3}$. Predictor LWCADJ splits internal node 3: leading to terminal node 3 (class 0) if LWCADJ $<$ 0.004 g m$^{-3}$; otherwise to one of four terminal nodes (two of which are class 1).

Regarding temperature and dewpoint, the smaller the difference between them, the lower the LCL and the more likely that the CHEF site will be in fog. CART, however, does not use them sequentially (see Fig. 7), so the physical interpretation is somewhat unclear; $T <$ 13.1°C and TD $< -1.9$°C lead independently to te minal nodes of class 1.

Regarding wind direction, which is used to split i ternal node 4, the mainly easterly sector (WD $<$ 185 leads to terminal node 4 (class 0); however, this des nation is only reached via internal node 3 (i.e., CMS $\geq$ 1408 m) and LWCADJ $\geq$ 0.004 g m$^{-3}$, so the phy ical interpretation is again difficult.

As shown in Table 7, the largest number of cas arriving at class 1 are those in terminal node 2, whi is reached by CMSL $<$ 1408 m and LWCEST $\geq$ 0.0 g m$^{-3}$. In summary, therefore, CART's selection a use of the split variables often makes physical sen post facto, which is somewhat reassuring. At oth times, the physical interpretation is somewhat obscu What is clear, however, is that it is very difficult determine a priori which variables are the most imp tant predictors and how they should be used in the cl sification tree to minimize prediction variance.

The results from CART run 1 are summarized in T bles 8a and 8b. Essentially, these consist of contingen tables for the learning and test sets separately (Tab 8a) and combined (Table 8b). Although the total a curacy values computed from (6) appear respectab (0.881–0.891), they are misleading, as they are dom nated by the large numbers in the (0, 0) boxes (top l of each table). More revealing is the fact that 2280 cas out of 2299 in the learning set, or 0.992, were predict correctly to fall in class 0 (no fog). Almost identi accuracy was achieved in the test set and in the co bined results (Table 8b). Predictions of class 1, howev had accuracy rates of only 0.422–0.440. (Similar resu are obtained from the more robust threat scores.) T results clearly demonstrate that the algorithm deriv from CART run 1 can be used with a high degree confidence to eliminate cases where fog was not o served to occur at Roundtop. The remaining data w contain a mixture of fog and no-fog events that can th be processed by a subsequent CART application.

### c. *Application of rules from CART step 1 run to 1988–91 dataset*

The prediction rules or algorithm described in Fig and Table 5 were used to process the original data ($N$ = 7357). As a result, 6022 cases, or 82%, we predicted to fall in class 0 (no fog). This compares clo ly to the 2299 cases out of 2854, or 81%, in the learni set of Table 8a and a corresponding 81% in the test of Table 8a. As mentioned above, these 6022 cases c be set aside with high confidence that they contain ca in which no fog was observed. The remaining 13 cases are expected to contain fog and no-fog cases, w the fog cases representing a larger proportion in t reduced dataset than they did in the complete datas

TABLE 8a. Contingency table for CART prediction of fog occurrence at Roundtop (class 0 = no fog; class 1 = fog). The learning set was a randomly selected 80% ($N$ = 2854) of the 1988–91 dataset ($N$ = 3567). The test set was the remaining 20% ($N$ = 713).

| | Learning set | | | | | Test set | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Predicted class | Observed class | | | | | Observed class | | | | |
| | 0 | 1 | Total | Accuracy | Threat | 0 | 1 | Total | Accuracy | Threat |
| 0 | 2280 | 19 | 2299 | 0.992 | 0.870 | 576 | 3 | 579 | 0.995 | 0.881 |
| 1 | 321 | 234 | 555 | 0.422 | 0.408 | 75 | 59 | 134 | 0.440 | 0.431 |
| Total | 2601 | 253 | 2854 | 0.881 | | 651 | 62 | 713 | 0.891 | |

TABLE 8b. As for Table 8a except the learning and test sets have been combined ($N$ = 3567).

| Predicted class | Observed class | | | | |
|---|---|---|---|---|---|
| | 0 | 1 | Total | Accuracy | Threat |
| 0 | 2856 | 22 | 2878 | 0.992 | 0.872 |
| 1 | 396 | 293 | 689 | 0.425 | 0.412 |
| Total | 3252 | 315 | 3567 | 0.883 | |

### *d. CART prediction of LWC*

The reduced dataset ($N$ = 1335) was processed by CART in a run referred to hereafter as run 2. Figure 8 shows the regression tree from this run. CART's tree-building process is described in section 1a; details of the splitting are given in Table 9. Terminal-node information is in Table 10 and is discussed below. Again, 20% of the data (266 cases) was reserved for testing purposes. A large number of predictors were employed for splitting the learning set data ($N$ = 1069). New predictors not previously defined are relative humidity (RH), the sine of the solar declination (SDECL), the cosine of JDSUM (CJDSUM), wind speed (WS), the sine of the solar altitude (SALT), an estimate of the average air density in the layer between the elevations of Sherbrooke and the Roundtop CHEF site (RHOBAR), the cosine of the wind direction relative to the Sherbrooke–Roundtop direction of 240° (CWD), and the Sherbrooke station pressure (PR).

The new predictors mentioned in the previous paragraph were, in fact, available at all previous stages of the CART applications. CART chose to ignore them before, but found it advantageous to use them in run 2. (As mentioned briefly in section 1a, the basis for CART's use of an available predictor is the minimization of variance between the predicted and observed data in the learning set.)

Table 10 presents the terminal node information for both the learning set and test set tests. As before, the predicted average values derived from the learning set should be compared with the observed average values in the test set. The standard deviations must also be considered when assessing the usefulness of the prediction. The results of Table 10 are presented graphically in Fig. 9. As before, the large standard deviations seem to be a problem. Average values are predicted well at some of the terminal nodes (e.g., 1, 7, 9, 12, 16, and 18) but poorly at others (e.g., 2, 3, 8, and 13).

The algorithm derived from CART run 2 was applied to the entire reduced dataset (i.e., the combined learning and test sets) to produce the comparison between calculated and observed LWC in Fig. 10. The correlation coefficient is 0.626, suggesting some degree of skill, but the scatter is rather large. Nevertheless, the predictors selected by CART run 2 are potentially useful to NFIS which is itself unable to eliminate predictors.

Figure 11 combines the no-fog results from CART run 1 with the CART run 2 results from Fig. 10. The correlation coefficient is 0.713, a significant improvement over the results of the one-step CART application shown in Fig. 6. Likewise, the slope and intercept of the linear regression line are improved.

Table 11 shows a contingency table that summarizes the combined results of CART runs 1 and 2. (The total of 6041 cases predicted in the 0–0.04 g m$^{-3}$ bin includes 6022 zero values from CART run 1. From the reduced dataset of 1335, therefore, CART run 2 predicted an additional 19 no-fog cases.) Comparison with the results of the one-step CART application in Table 5 shows that the no-fog and fog prediction accuracies are only slightly improved (0.952 to 0.965 and 0.498 to 0.510, respectively). The threat score in the no-fog category is only slightly improved, whereas in the fog category it increased from 0.398 to 0.440. Although the improvement in the accuracy of fog prediction (i.e., absence or presence) is modest, there has been a very significant

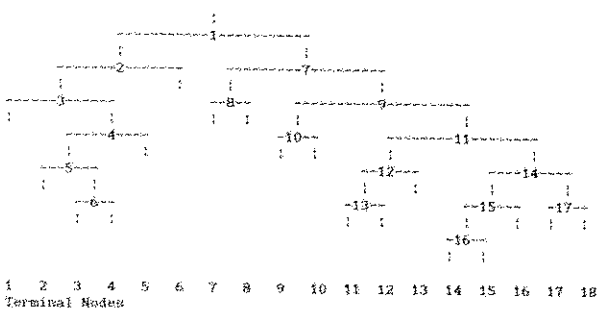


FIG. 8. CART regression tree diagram for LWC prediction from the reduced dataset ($N$ = 1335). The 17 split variables (one for each internal node) are listed in Table 9. Terminal node information appears in Table 10.

TABLE 9. Learning set node information for LWC prediction from reduced dataset. Terminal nodes are shown in bold figures. The tr structure is shown in Fig. 8.

| Node | Cases | Split variable | Threshold | Units | Left node | Right node |
|---|---|---|---|---|---|---|
| 1 | 1069 | RH | 91.45 | % | 2 | 7 |
| 2 | 586 | CMSL | 793 | m | 3 | **6** |
| 3 | 212 | SDECL | −0.25 | — | **1** | 4 |
| 4 | 211 | CJDSUM | 0.997 | — | 5 | **5** |
| 5 | 200 | WS | 2.2 | m s$^{-1}$ | **2** | 6 |
| 6 | 99 | SALT | 0.444 | — | **3** | **4** |
| 7 | 483 | JDSUM | −18.5 | day | 8 | 9 |
| 8 | 61 | RHOBAR | 1.148 | kg m$^{-3}$ | **7** | **8** |
| 9 | 422 | CWD | 0.421 | — | 10 | 11 |
| 10 | 138 | SDECL | 0.396 | — | **9** | **10** |
| 11 | 284 | JDSUM | 71.5 | day | 12 | 14 |
| 12 | 148 | RHOBAR | 1.141 | kg m$^{-3}$ | 13 | **13** |
| 13 | 74 | T | 18.85 | °C | **11** | **12** |
| 14 | 136 | SDECL | −0.073 | — | 15 | 17 |
| 15 | 70 | SDECL | −0.101 | — | 16 | **16** |
| 16 | 56 | PR | 982.9 | hPa | **14** | **15** |
| 17 | 66 | TD | 8.25 | °C | **17** | **18** |

improvement in the accuracy of the LWC prediction, as shown in the paragraph above.

### *e. NFIS prediction of LWC*

Using the predictors from CART run 2 and applying them to the reduced dataset ($N$ = 1335) produced a prediction algorithm that was used to calculate values of LWC. These 1335 values were then combined with the 6022 values set aside by application of the CART run 1 algorithm, all of which had predicted values of 0 g m$^{-3}$. The complete prediction of 7357 cases is compared with the observed values in Fig. 12. The Pearson linear correlation coefficient is a respectable $r$ = 0.810, a significant improvement over the results ($r$ = 0.596) presented in Fig. 6. The best-fit linear regression l has a slope of 0.671 and a $y$ intercept of 0.003, indicat a tendency to underpredict. Nevertheless, about 10 the 11 highest predicted values lie close to the 1:1 p fect-fit line.

Table 12 presents the contingency table resulting fr the three-step (CART–CART–NFIS) application. ( total of 6398 cases predicted in the 0–0.04 g m$^{-3}$ includes 6022 zero values from CART run 1. From reduced dataset of 1335, therefore, NFIS predicted additional 376 no-fog cases.) The accuracy of the fog prediction is essentially unchanged from the tv step (CART–CART) process. The accuracy of the prediction, on the other hand, underwent a signific

TABLE 10. Terminal node information for LWC (g m$^{-3}$) prediction from reduced dataset.

| | Learning set | | | Test set | | |
|---|---|---|---|---|---|---|
| Node | Cases | Avg | Std dev | Cases | Avg | Std dev |
| 1 | 1 | 0.538 | 0.005 | 1 | 0.521 | 0.000 |
| 2 | 101 | 0.070 | 0.086 | 27 | 0.108 | 0.093 |
| 3 | 46 | 0.071 | 0.084 | 10 | 0.178 | 0.141 |
| 4 | 53 | 0.180 | 0.126 | 16 | 0.219 | 0.135 |
| 5 | 11 | 0.232 | 0.170 | 2 | 0.188 | 0.188 |
| 6 | 374 | 0.044 | 0.079 | 102 | 0.063 | 0.099 |
| 7 | 36 | 0.403 | 0.119 | 8 | 0.437 | 0.090 |
| 8 | 25 | 0.175 | 0.180 | 3 | 0.106 | 0.149 |
| 9 | 117 | 0.053 | 0.083 | 33 | 0.072 | 0.118 |
| 10 | 21 | 0.226 | 0.135 | 6 | 0.184 | 0.137 |
| 11 | 29 | 0.238 | 0.096 | 7 | 0.161 | 0.131 |
| 12 | 45 | 0.115 | 0.099 | 10 | 0.116 | 0.118 |
| 13 | 74 | 0.090 | 0.088 | 15 | 0.131 | 0.097 |
| 14 | 26 | 0.251 | 0.118 | 7 | 0.303 | 0.152 |
| 15 | 30 | 0.095 | 0.100 | 4 | 0.123 | 0.146 |
| 16 | 14 | 0.003 | 0.011 | 5 | 0.000 | 0.000 |
| 17 | 19 | 0.126 | 0.111 | 3 | 0.109 | 0.082 |
| 18 | 47 | 0.287 | 0.092 | 7 | 0.263 | 0.109 |
| Total | 1069 | | | 266 | | |

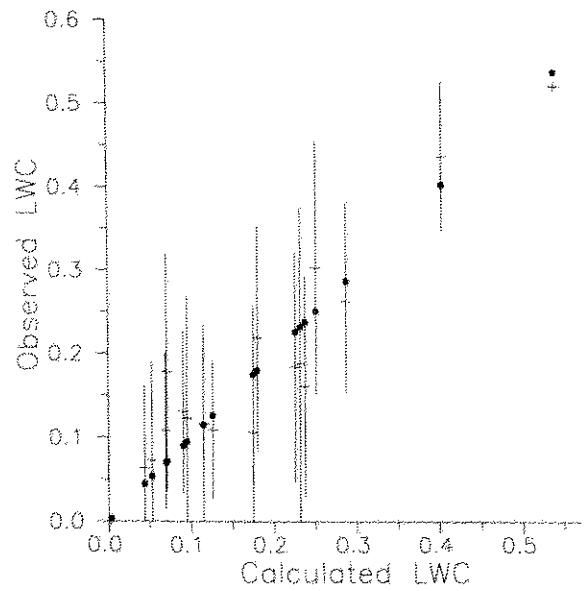


FIG. 9. As in Fig. 5 except the test set ($N$ = 266), representi 20% of the reduced dataset ($N$ = 1335), is plotted (see Table

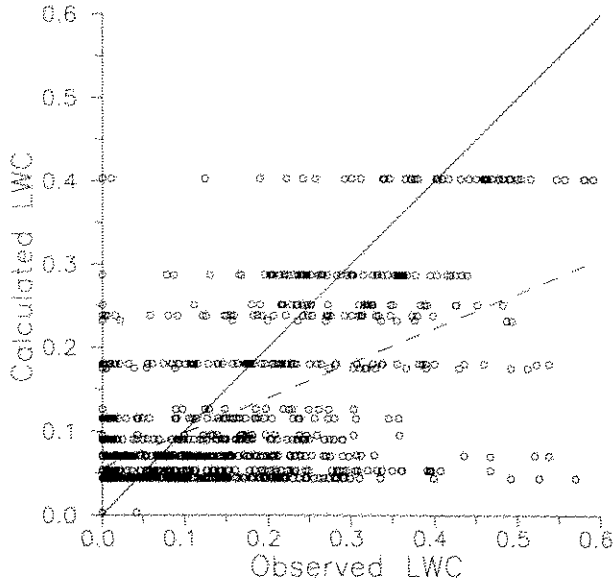


FIG. 10. As in Fig. 6 except the reduced 1988–91 dataset ($N$ = 1335) is plotted; $r$ = 0.626, $a_1$ = 0.418, and $a_0$ = 0.056 g m$^{-3}$.

improvement from 0.510 to 0.664. Furthermore, the more robust threat score improved from 0.872 to 0.916 in the no-fog category and from 0.440 to 0.530 in the fog category.

## 7. Summary and conclusions

The classic problem of accurate prediction of an infrequent event was encountered in this project. As shown in section 3, fog was observed by the OCD at the Roundtop CHEF site only about 15% of the time among 4093 hourly records in the spring, summer, and autumn of 1988–91. (When hours in which fog occurred for part of the time were eliminated, the frequency of complete hours of fog was reduced to 10%.) Nevertheless, a methodology was developed that achieved the objective of this paper and yielded acceptable estimations of the LWC in high-elevation fog on Roundtop Mountain from routine meteorological measurements at the Sherbrooke weather station.

TABLE 11. Contingency table for the combined CART–CART prediction of LWC (g m$^{-3}$) at Roundtop. The linear correlation coefficient is $r$ = 0.713.

| Predicted LWC | Observed LWC | | | | |
|---|---|---|---|---|---|
| | 0–0.04 | >0.04 | Total | Accuracy | Threat |
| 0–0.04 | 5832 | 209 | 6041 | 0.965 | 0.872 |
| >0.04 | 645 | 671 | 1316 | 0.510 | 0.440 |
| Total | 6477 | 880 | 7357 | 0.884 | |

Table 13 presents a summary of the statistical measures used to quantify the skill. The bottom row of Table 13 shows that the CART–CART–NFIS method achieved a correlation coefficient of 0.810 and threat scores of 0.916 and 0.530 for the no-fog and fog events, respectively. All three of these measures of skill are significantly superior to those of earlier methods, even those that used CART. (Accuracies corresponding to the above threat scores were 0.962 and 0.664, respectively.)

Figure 13 summarizes the three-step procedure adopted to circumvent the problem of predicting an infrequent event. CART run 1 was designed to work on the FD data in a classification mode to predict the absence of

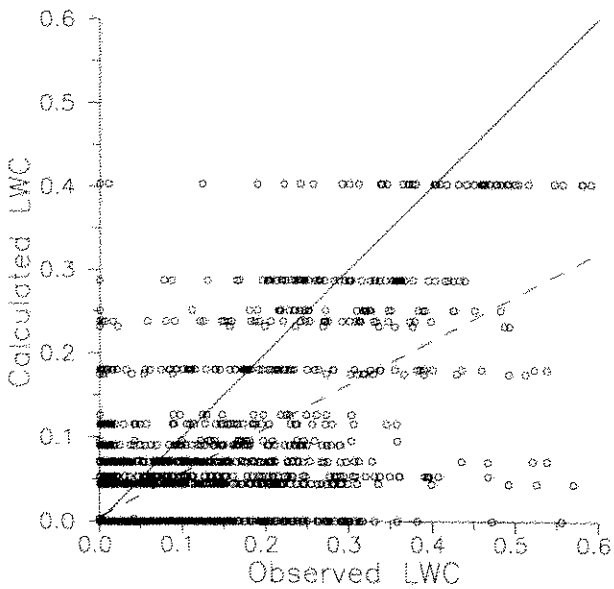


FIG. 11. As in Fig. 10 except CART results from run 1 ($N$ = 6022) are combined with CART results for the reduced dataset ($N$ = 1335). Combined $N$ = 7357; $r$ = 0.713, $a_1$ = 0.526, and $a_0$ = 0.005 g m$^{-3}$.

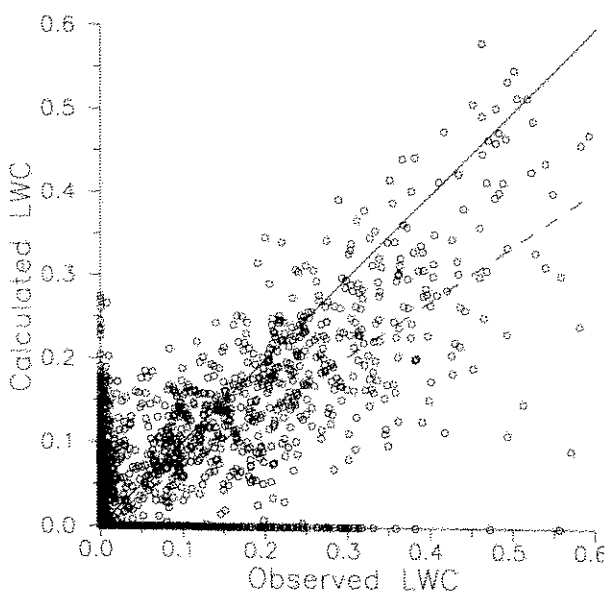


FIG. 12. As in Fig. 11 except CART results from run 1 ($N$ = 6022) are combined with NFIS results for the reduced dataset ($N$ = 1335). Combined $N$ = 7357; $r$ = 0.810, $a_1$ = 0.671, and $a_0$ = 0.003 g m$^{-3}$.

TABLE 12. Contingency table for the combined CART–NFIS prediction of LWC (g m$^{-3}$) at Roundtop. The linear correlation coefficient is $r = 0.810$.

| Predicted LWC | Observed LWC | | | | |
|---|---|---|---|---|---|
| | 0–0.04 | >0.04 | Total | Accuracy | Threat |
| 0–0.04 | 6155 | 243 | 6398 | 0.962 | 0.916 |
| >0.04 | 322 | 637 | 959 | 0.664 | 0.530 |
| Total | 6477 | 880 | 7357 | 0.923 | |

presence of fog. Referring to Fig. 13a, note that in 2878 out of 3567 cases, or 81% of the time, CART predicted that fog did not occur. The accuracy rate of this no-fog prediction was 0.992. The accuracy rate of the CART run 1 fog prediction, on the other hand, was only 0.425. Threat scores for the no-fog and fog predictions were 0.872 and 0.412, respectively.

The algorithm derived from CART run 1 was applied to all the 7357 hourly records for which LWC was available (either as a measured value or a zero value derived from the FD measurement), as shown in Fig. 13b. Fog was not predicted to occur in 6022 cases, or 82% of the time. Those cases were set aside with high confidence that they, indeed, were no-fog events. The remaining 1335 cases were processed by CART run 2, which was designed to predict the actual LWC values in a regression mode. The algorithm from this run, when applied to the 1335 cases and compared with observed values, produced a Pearson linear correlation coefficient of 0.626. Although this result was not spectacular, it did suggest which predictors were significant. As NFIS is unable to eliminate predictors, prescreening by CART run 2 proved very beneficial. NFIS was able to work efficiently in step 3 of this methodology using the predictors suggested by CART. Comparison of the LWC calculated by the CART run 1 and NFIS algorithms with the complete set of observed and derived LWC yielded a correlation coefficient of 0.810. Accuracy rates for no-fog and fog prediction were 0.962 and 0.664, respectively; corresponding threat scores were 0.916 and 0.530.

Despite some remaining scatter in the final result (Fig. 12), the correlation between calculated and observed LWC values is greatly superior to the initial prediction attempts made without the aid of CART or NFIS (Fig. 2) and even the secondary attempts made with CART applications to the complete dataset (Fig. 6). When "infrequent event" problem was fully appreciated, idea of screening out the predictions of the "frequ event" (the nonoccurrence of fog in this proje emerged. It should be stressed that we used the CA prediction of no fog, not the observations, to perfo this screening.

We feel confident that the algorithms developed h are applicable to prediction at the Roundtop CHEF in the spring, summer, and fall seasons of years outs the 1988–91 period for which verification data exist should be reiterated that the FD and LWC data are needed for applying the CART and NFIS algorithm although if the data are available, they are useful verification.

What other measures could we have taken to m our task easier? We could have used, for example, meteorological classifications of Brook et al. (1995) an additional predictor. In effect, this would have b a means of allowing CART to use a preliminary matological screening of the data; however, there is guarantee that CART would have found that this p dictor helped to reduce variance. (It should be no that precipitation type and intensity were available predictors, but CART evidently did not find that t helped to reduce variance at any stage of the proc so it did not use them.) Another possibility would h been to run back trajectories from Roundtop to de mine the source of the air. Both of these ideas wo require a considerable effort to implement. In the we opted to keep the list of predictors as simple possible in the hope that we could develop a sche that would rely only on routinely available data hence have a possibility of being more universally plicable.

Still another source of predictors would be objectiv analyzed, gridded upper-air data interpolated to Roundtop (or Sherbrooke weather station) location. T would be a more routine source than either the m orological classifications or the trajectory analy however, there would still be some effort involve merging the upper-air and surface data so that CA could process them. Furthermore, there would agai no guarantee that CART would select any of the up air predictors.

The remaining challenges in this project will b

TABLE 13. Summary of statistical evaluation.

| Period | $N$ | Method | Correlation | LWC ≤ 0.04 | | LWC > 0.04 | |
|---|---|---|---|---|---|---|---|
| | | | | Accuracy | Threat | Accuracy | Threat |
| 1988 | 2886 | First estimate | 0.321 | | | | |
| 1988 | 2886 | Adjusted estimate | 0.479 | | | | |
| 1988 | 2886 | Smoothed | 0.531 | | | | |
| 1988–91 | 7357 | CART | 0.596 | 0.952 | 0.869 | 0.498 | 0.398 |
| 1988–91 | 7357 | CART–CART | 0.713 | 0.965 | 0.872 | 0.510 | 0.440 |
| 1988–91 | 7357 | CART–CART–NFIS | 0.810 | 0.962 | 0.916 | 0.664 | 0.530 |

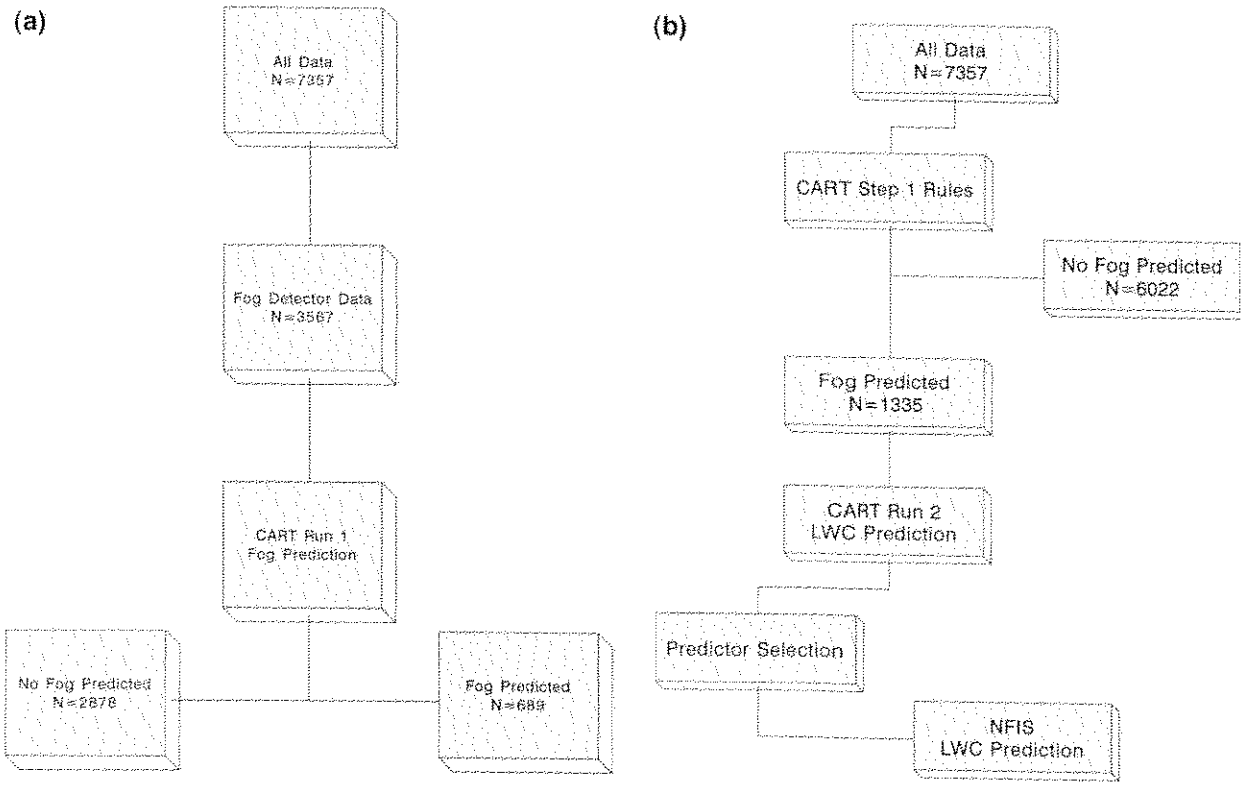


FIG. 13. LWC prediction scheme. (a) Step 1 and (b) steps 2 and 3.

to extend the method to apply to other elevations at Roundtop and hence enable an estimation of spatial variability following methods outlined in Walmsley et al. (1996a) and 2) to apply the same or similar algorithms at other sites for which verification data are available in an effort to improve the universality of the method. As previously mentioned, we have tried to minimize the uncertainty in extending the method to other locations by not using data specific to Roundtop, except for the LWC measurements used to develop the algorithms and perform the verification; that is, no Roundtop data are used as predictors. One of the main problems, however, would be in finding additional data for training and verification purposes. Apart from CHEF, two sources that could be considered are the MCCP and the Great Dun Fell Experiment (Choularton et al. 1997).

*Acknowledgments.* Thanks to Richard Tanabe for extracting the hourly observed LWC and FD data from the CHEF database, to Mohammed Wasey for information about the fog measuring device, and to Dr. Yi-Fan Li for Fig. 1. Dr. Peter Summers drew our attention to the problem of predicting infrequent events highlighted in the *New Scientist* article of Matthews (1997). Prof. Douw Steyn asked two very relevant questions at a conference presentation of this material, which have served to focus our discussion. Howard Barker completed a thorough review of the manuscript; his helpful comments are gratefully acknowledged. The two reviewers provided excellent and thoughtful comments The CART software is marketed by Salford Systems 8880 Rio San Diego Drive, Suite 1045, San Diego, CA 92108. The NFIS software is marketed by The Mat Works, Inc., 24 Prime Park Way, Natick, MA 01760 1500.

## REFERENCES

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

Bridgman, H. A., J. L. Walmsley, and R. S. Schemenauer, 1994 Modelling the spatial variations of wind speed and direction o Roundtop Mountain, Quebec. *Atmos.–Ocean,* **32,** 605–619.

Brook, J. R., P. J. Samson, and S. Sillman, 1995: Aggregation o selected three-day periods to estimate annual and seasonal we deposition totals for sulfate, nitrate, and acidity. Part I: A syn optic and chemical climatology for eastern North America. *J Appl. Meteor.,* **34,** 297–325.

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

Chiu, S., 1994: Fuzzy model identification based on cluster estima tion. *J. Intell. Fuzzy Syst.,* **2,** 269–278.

Choularton, T. W., and Coauthors, 1997: The Great Dun Fell Clou Experiment 1993: An overview. *Atmos. Environ.,* **31,** 2193 2405.

Gerber, H., 1984: Liquid water content of fogs and hazes from visibl light scattering. *J. Climate Appl. Meteor.,* **23,** 1247–1252.

Mathworks, 1998: Fuzzy logic toolbox for Use with MATLAB: User' guide. The Mathworks Inc., Natick, MA, 231 pp.

Matthews, R., 1997: How right can you be? *New Sci.,* **153,** 28–31.

Schemenauer, R. S., C. M. Banic, and N. Urquizo, 1995: High elevation fog and precipitation chemistry in southern Quebec, Canada. *Atmos. Environ.*, **29,** 2235–2252.

Stanski, H. S., L. J. Wilson, and W. R. Burrows, 1990: Survey of common verification methods in meteorology. World Meteorological Organization Tech. Doc. WMO/TD No. 358, 114 pp.

Steinberg, D., and P. Colla, 1995: *CART: A Supplementary Module for SYSDAT.* Salford Systems, 307 pp.

Urquizo, N., J. L. Walmsley, W. R. Burrows, R. S. Schemenauer, and J. R. Brook, 1998: Application of the CART and NFIS statistical analyses to fogwater and the deposition of wet sulphate in mountainous terrain. *Air Pollution Modeling and Its Application,* Vol. 12, S.-E. Gryning and N. Chaumerliac, Eds., Plenum Press, 409–417.

Walmsley, J. L., H. A. Bridgman, and R. S. Schemenauer, 19 Modelling fog water deposition of sulfate on Roundtop Mount. Quebec. *Proc. Int. Conf. on Modelling and Simulation,* Vol Newcastle, NSW, Australia, Modelling and Simulation Soc. of Australia, Inc., 67–71.

——, R. S. Schemenauer, and H. A. Bridgman, 1996a: A method estimating the hydrologic input from fog in mountainous terr *J. Appl. Meteor.,* **35,** 2237–2249.

——, N. Urquizo, R. S. Schemenauer, and H. A. Bridgman, 19 Modelling of acid deposition in high-elevation fog. *Air Pollu. IV. Monitoring, Simulation and Control,* B. Caussade, H. Po and C. A. Brebbia, Eds., Computational Mechanics Publicati 195–206.