Sonoma Technology, Inc.

1360 Redwood Way, Suite C
Petaluma, CA 94954-1169
707/665-9900
FAX 707/665-9800
www.sonomatech.com

October 22, 2003

Michael Koerber
Lake Michigan Air Directors Consortium
2250 East Devon Ave. Suite 250
Des Plaines, IL 60018

STI Ref. No. 903480

Re: $PM_{2.5}$ Forecasting – Statistical Forecasting Tool Development

Dear Mike,

Thank you for giving Sonoma Technology Inc. (STI) the opportunity to provide data analysis and $PM_{2.5}$ forecasting services to six cities in the Midwest. As part of this effort, this technical memorandum summarizes the work developing statistical tools for forecasting summertime $PM_{2.5}$ concentrations in the six Midwest cities. The following tasks were performed to complete this work:

- Created a database combining daily $PM_{2.5}$ concentrations from Federal Reference Method (FRM) monitors and meteorological variables from the National Weather Service (NWS) surface and upper-air data for each city during summer months, May–September, 1999-2002.
- Brought data into a statistical package and developed next-day forecasting tools using linear regression equations or Classification and Regression Trees (CART).
- Tested tools on either 2003 summer data, when available, or a subset of data from 1999-2002 that were not used in tool development.

Our findings are summarized in this letter with three attachments: Attachment 1 contains a more detailed discussion of results and methods, Attachment 2 contains variable definitions, and Attachment 3 contains references.

Results

CARTs were developed for the summer (May–September) for each of the six Midwestern cities: Chicago, Milwaukee, Cleveland, Columbus, Detroit and St. Louis. **Table 1** shows the explanations of the variance (r^2) for each CART and the accuracy (percent of time the AQI category was correctly predicted) of each CART for each city.

Following are the key findings of this effort:

- The explanation of the variance (r^2) was 0.61 or above for all cities except St. Louis, for which the value was 0.56. Generally, the goal for regression trees is an r^2 value above 0.60.
- Using test data sets, the regression trees were correct in predicting the AQI category between 67% of the time at Cleveland and Milwaukee and 86% of the time at Columbus. May through August 2003 data for all cities except Milwaukee were used to test results. Milwaukee $PM_{2.5}$ data were not readily available for 2003, so approximately 15% of the 1999-2002 data were reserved for testing (i.e., the data were not used in development of the regression trees).
- Overall, the regression trees developed in this work will provide a useful tool in assisting stakeholders in forecasting summertime $PM_{2.5}$ concentrations.

Table 1. Results of regression trees for each city: explanation of the variance (r^2) on development data set and accuracy (% of time AQI category was predicted correctly on test data set).

City	r^2	Accuracy (%)	Number of samples in test data
Chicago	0.64	79	78
Milwaukee	0.69	67	64
Cleveland	0.63	65	63
Columbus	0.67	86	77
Detroit	0.61	66	74
St. Louis	0.56	67	72

The following attachments provide details supporting our conclusions: Attachment 1, Development of Classification and Regression Trees; Attachment 2, Variable Definitions; and Attachment 3, References. If you have any questions about this memorandum, please contact either Tim Dye or me.

Sincerely,

Steven G. Brown
Air Quality Analyst

Timothy S. Dye
Vice President Meteorological Programs and Public Outreach

ATTACHMENT 1: DEVELOPMENT OF CLASSIFICATION AND REGRESSION TREES

CART OVERVIEW

Classification and Regression Tree (CART) is a statistical procedure that classifies data points into separate groups. In terms of $PM_{2.5}$ forecasting, CART allows for development of a decision tree to assist forecasters in predicting $PM_{2.5}$ concentrations based on values of meteorological variables output by weather models.

CART utilizes a data set of daily maximum $PM_{2.5}$ concentrations and meteorological variables to split the maximum daily $PM_{2.5}$ concentrations into separate groups based on a value of a meteorological variable (National Research Council, 1991; Stoeckenius, 1990). This threshold cutoff value is statistically determined so that the split most effectively separates days of different $PM_{2.5}$ concentration ranges. Variables selected by CART are generally those that have a high correlation with $PM_{2.5}$ concentrations. The resultant decision trees can then be used by forecasters as an aid to estimate the likely Air Quality Index (AQI) category using forecasted meteorological data. This procedure can be either manual or automated with a database.

The strengths of the CART tool (U.S. Environmental Protection Agency, 2003) are that it

- requires little expertise to operate,
- can be run quickly,
- complements subjective forecasting methods, and
- allows differentiation between days with similar $PM_{2.5}$ concentrations if $PM_{2.5}$ levels are a result of different meteorological processes.

CART also has some limitations:

- CART development requires expertise and effort.
- Small changes in predictor variables may lead to significant differences in the predicted $PM_{2.5}$ value.
- CART is an objective tool, so it is only as good as the available data. If the meteorological forecast does not verify, neither will the CART because the values of the predictor variables from the forecast were incorrect.
- CART will generally not characterize well unusual patterns or emissions (such as the high $PM_{2.5}$ concentrations often seen on the Fourth of July), although human forecasters can often account for these changes and use their discretion in applying the procedure.

DATA USED IN CART DEVELOPMENT

Twenty-four-hour $PM_{2.5}$ concentrations are measured at multiple sites in each city; the maximum concentration from all sites in a city was used as the maximum for the day. The $PM_{2.5}$ data were quality-controlled (QC'd), so that if the maximum $PM_{2.5}$ concentration at a single site was significantly higher than at most other sites, it was not used; and the next highest concentration among the remaining sites was used as a regionally representative concentration for the area. This screening was used to eliminate any localized, rare-event concentrations that may have affected the development of the CART trees.

Surface and upper-air meteorological variables were obtained from the NWS network. These data were also QC'd. Variable definitions are given in Attachment 2. Upper-air variables were taken from heights of 925 mb, 850 mb, 700 mb, and 500 mb and included temperature, height, and wind speed. Surface variables include 6-hr integrated wind speed and wind direction; temperature, dew point, relative humidity, pressure, wind speed, and wind direction every 12 hours; and precipitation, cloudiness, change in surface pressure, and maximum and minimum temperature every 24 hours. Additional parameters were calculated: 24-hr difference in 500-mb height; 24-hr and 12-hr difference in 925-mb height; and 12-hr differences in temperature between the surface and both 925-mb and 850-mb heights.

The previous day's $PM_{2.5}$ concentration for the city was also used. This value is actually a two-day difference, since the forecast is for "tomorrow's" $PM_{2.5}$. This value will be readily available to forecasters, unlike "today's" $PM_{2.5}$ concentration, which would not have occurred at the time of forecasting for tomorrow's $PM_{2.5}$. Additionally, previous day's $PM_{2.5}$ concentrations from cities that showed a high correlation in $PM_{2.5}$ concentrations (such as Cleveland and Columbus) were also used in CART development. These additions may help characterize transport and persistence, and increase the accuracy of the tool for Milwaukee and Detroit.

CART RESULTS

Multiple iterations using different variables and statistical restraints were conducted in developing the trees. Primarily, the goal was to maximize the r^2 value; however, if some cutpoints did not make physical sense (i.e., a lower upper-level height and higher $PM_{2.5}$ concentration), the tree was generally not used. This is a key concept because some variables may have a correlation with $PM_{2.5}$ under certain conditions, but if the results do not make physical sense, they should be scrutinized. Additionally, splits that conform to AQI categories are most useful, because the AQI category is the necessary result of any forecast. Thus, iteration also included optimizing the end nodes of the tree to achieve a split that best characterized the $PM_{2.5}$ concentrations by AQI category. These iterations are essential to maximize the accuracy of a given tree.

Final CART results for each city are shown in **Figures 1-1 through 1-6** and are summarized in **Table 1-1**. Over 300 samples were used for each city, except Milwaukee, where approximately 15% of the data were retained for testing and not used in tool development. The explanation of the variance (r^2) ranged between 0.61 and 0.69 for all cities except St. Louis, which had a value of 0.56.

Despite numerous iterations and adjustments to tolerances, the value for St. Louis could not be increased. Synoptic typing indicates that St. Louis may often have conditions that should encourage high $PM_{2.5}$ concentrations (i.e., stagnant air at the surface and aloft, a stationary front or surface high, etc.), but maximum concentrations are often only in the Moderate range. It may be that the atmospheric chemistry in the area is not well-characterized by meteorological variables alone, which would reduce the r^2 value. The lower r^2 for the summer months in this analysis is similar to results found for winter months as well (MacDonald et al., 2003).

At each city, the temperature at 925 mb or 850 mb provided the first cut, with lower temperatures yielding lower $PM_{2.5}$ concentrations, which is consistent with $PM_{2.5}$ formation. Surface dew point temperatures provided secondary and tertiary cuts for all cities except Cleveland; lower dew points were associated with lower $PM_{2.5}$. This is consistent with the idea of $PM_{2.5}$ formation by gas-to-particle conversion enhanced by cloud processing, especially for sulfate (Seinfeld and Pandis, 1998), which is often more than half the $PM_{2.5}$ in the Midwest. The previous day's $PM_{2.5}$ maximum concentration was used for Milwaukee, with a previously lower $PM_{2.5}$ concentration indicating lower $PM_{2.5}$ the next day. The previous day's $PM_{2.5}$ maximum at Milwaukee was used for Detroit also and is consistent with earlier analysis demonstrating a relationship between the two cities, likely due to transport (Brown and Dye, 2003). Wind speed was used in nearly all trees, since lower wind speed leads to higher $PM_{2.5}$ concentrations; this is consistent with stagnant conditions allowing $PM_{2.5}$ concentrations to increase. Other variables include wind direction (generally winds from the north and west were associated with lower $PM_{2.5}$ than winds from the south), upper-air height (lower heights were associated with lower $PM_{2.5}$), changes in surface pressure (a large rise in surface pressure associated with a high were associated with higher $PM_{2.5}$), and height differences between 925 mb and the surface (large differences indicate enhanced subsidence and would lead to higher $PM_{2.5}$ concentrations).

Table 1-1. Number of sites for $PM_{2.5}$ data, number of samples, explanation of the variance (r^2), and number of end nodes for each city's CART.

City	N monitoring sites	N samples for development	r^2	N end nodes
Chicago	26	348	0.64	7
Milwaukee	10	281	0.69	6
Cleveland	12	375	0.63	6
Columbus	5	388	0.67	8
Detroit	12	335	0.61	9
St. Louis	15	391	0.56	11

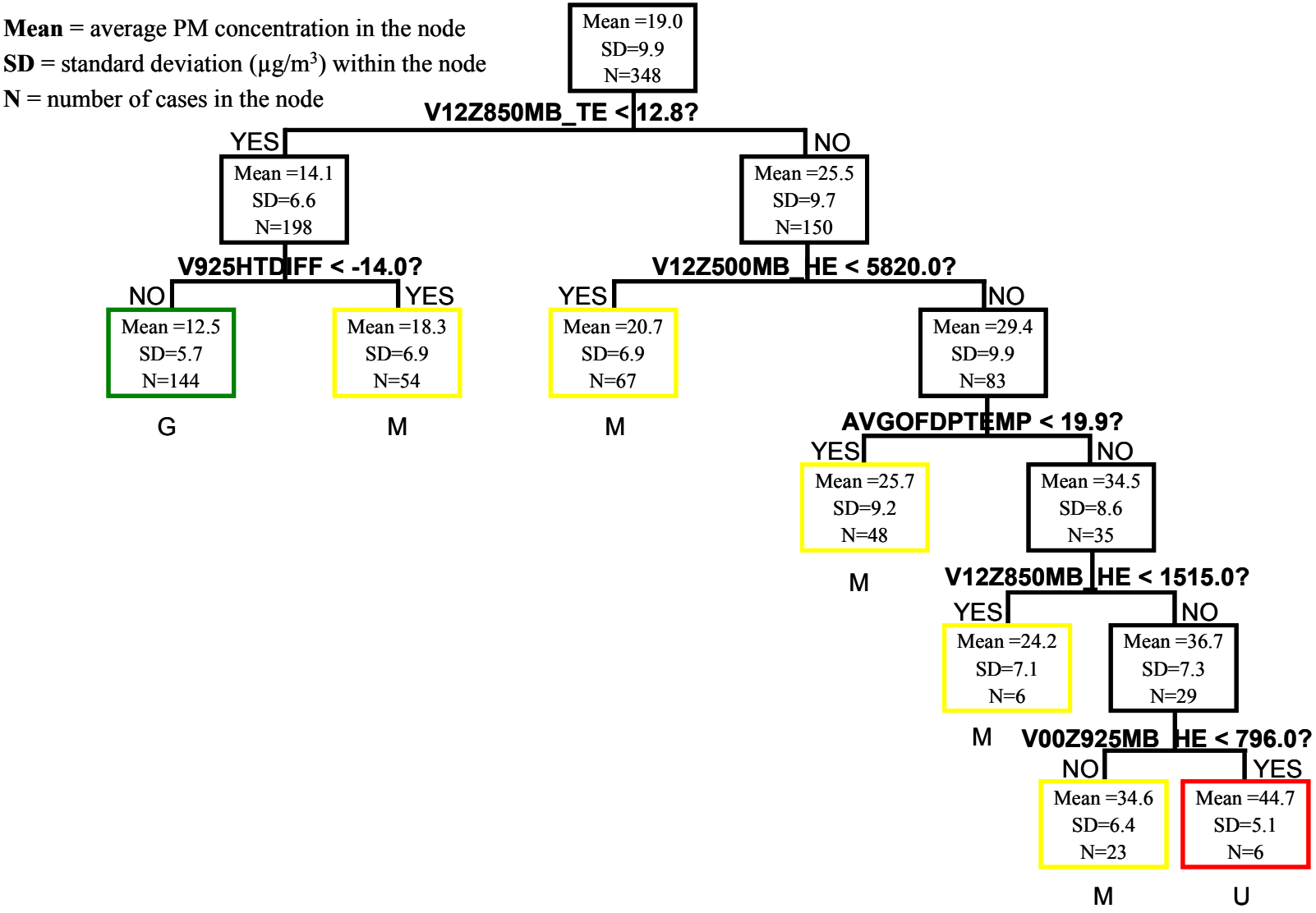

Figure 1-1. CART for Chicago.

Mean = average PM concentration in the node
SD = standard deviation (µg/m^3) within the node
N = number of cases in the node

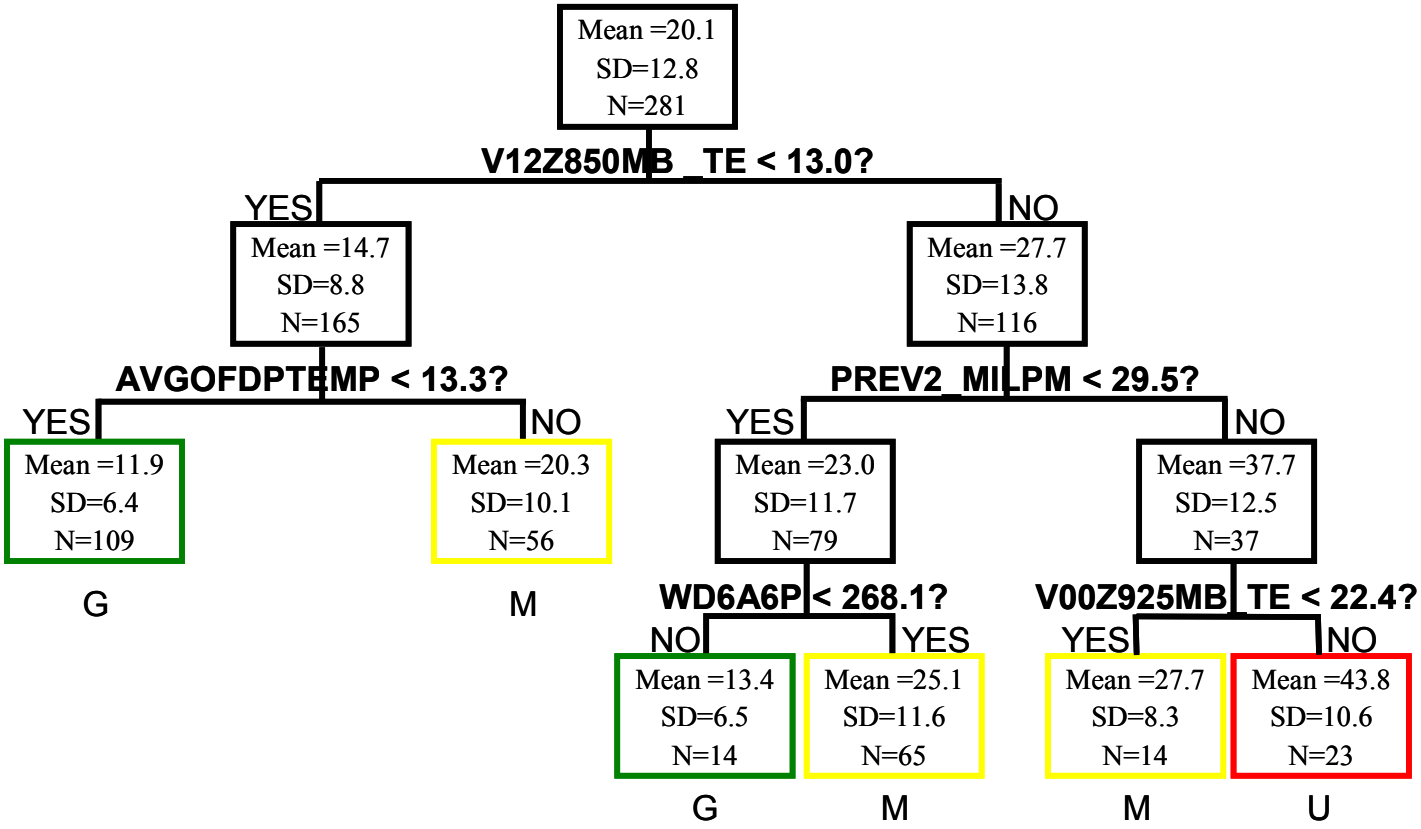

Figure 1-2. CART for Milwaukee.

Mean = average PM concentration in the node
SD = standard deviation (µg/m^3) within the node
N = number of cases in the node

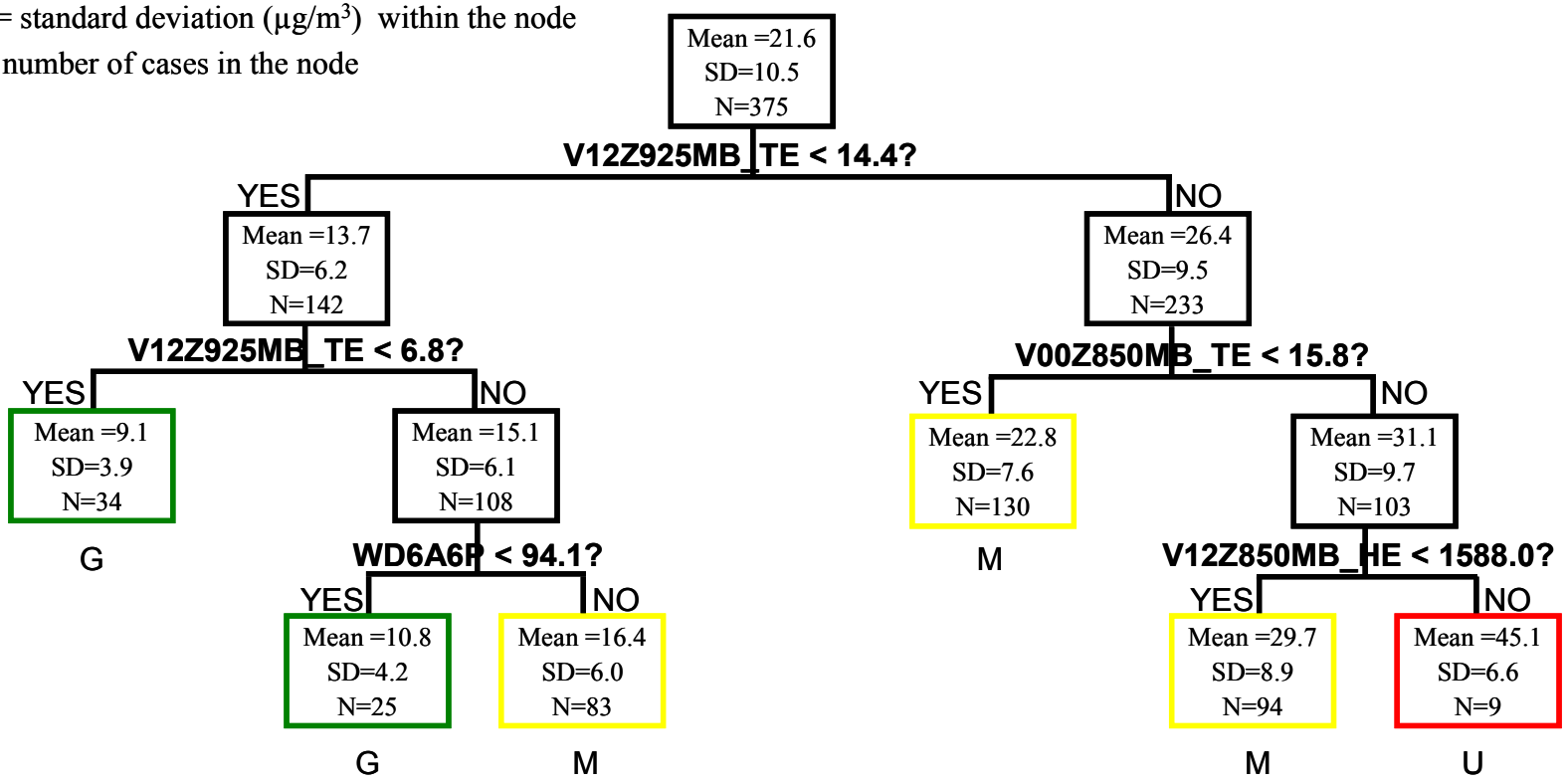

Figure 1-3. CART for Cleveland.

Mean = average PM concentration in the node
SD = standard deviation (μg/m^3) within the node
N = number of cases in the node

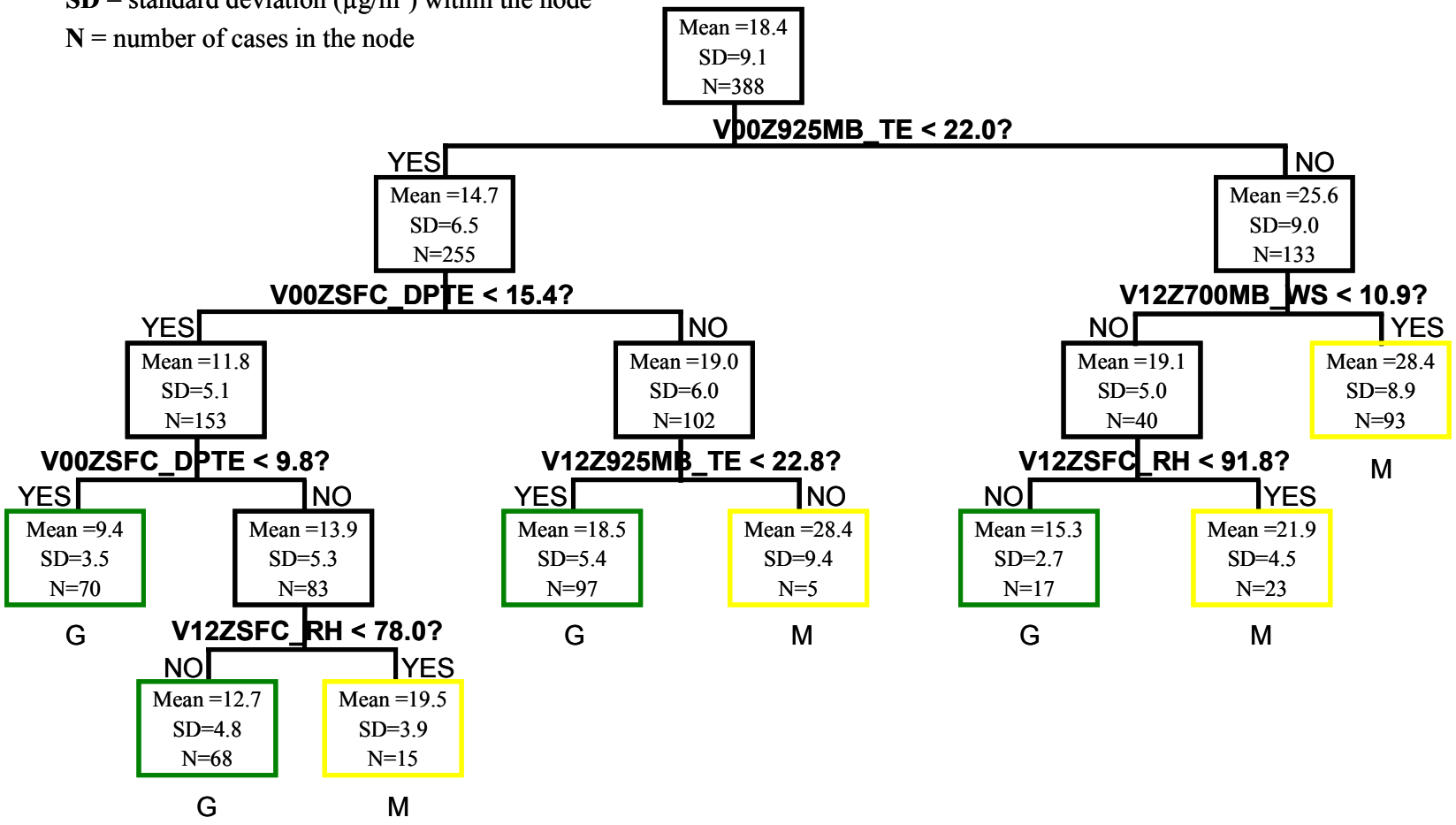

Figure 1-4. CART for Columbus.

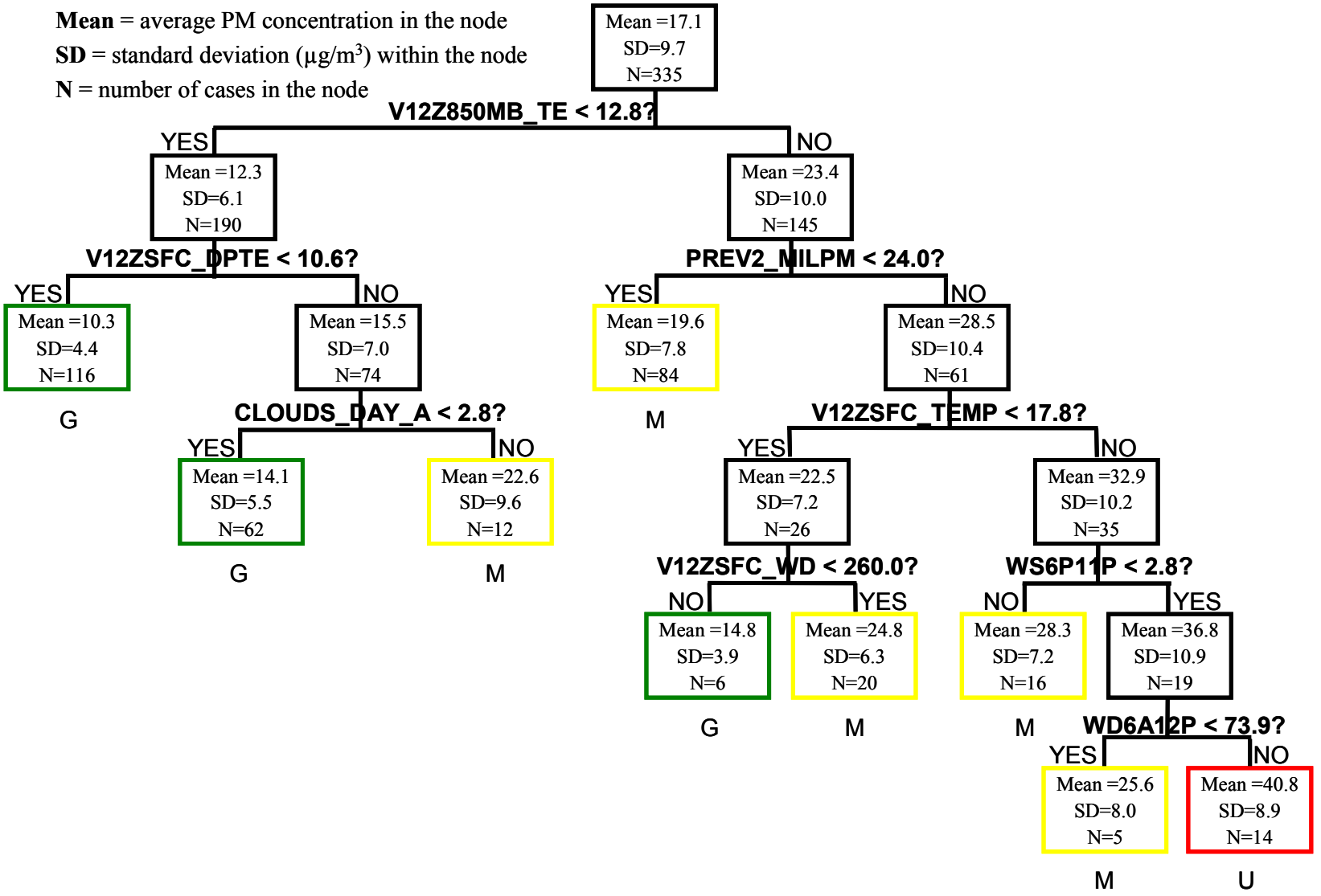

Figure 1-5. CART for Detroit.

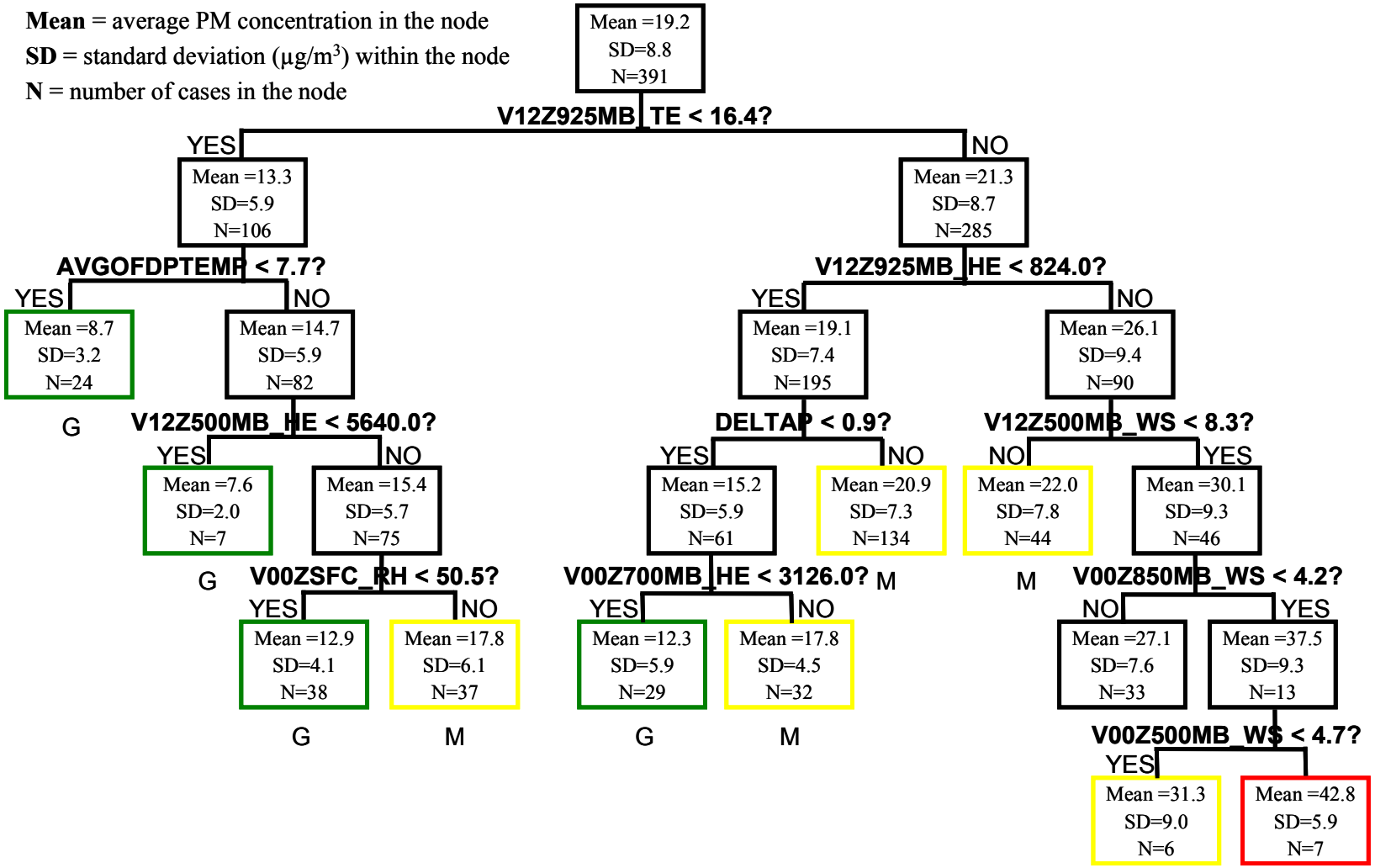

Figure 1-6. CART for St. Louis.

CART TESTING

While the explanation of the variance (r^2) is a good indicator of the effectiveness of a CART to predict $PM_{2.5}$ levels, testing it is a much better gauge of how it will perform operationally. Each CART was tested against summer (May–August) 2003 $PM_{2.5}$ data for all cities except Milwaukee. Milwaukee $PM_{2.5}$ data were not readily available, so a subset of 1999-2002 data was used; this subset was not included in the development of the CART, since testing the tree on data used to develop it would bias the results.

Test results for each city are shown in **Tables 1-2 through 1-7**. For each CART, the results for all test data and categories are given. These results show the number of correctly predicted values by AQI category, and the number of values incorrectly predicted. Overall, the CARTs were correct between 67% of the time at Cleveland and Milwaukee and 86% of the time at Columbus, with sample sizes between 63 and 78 samples. These results are encouraging and support the use of CART in assisting forecasters.

At Chicago, most of the daily maximums fell into the Moderate category and were predicted correctly about 75% of the time. About one-fourth of the values were in the Good category, and 83% of these were predicted correctly. No instances of Unhealthy for Sensitive Groups (USG) occurred.

For Cleveland, most of the daily maximums were also Moderate, and these were captured very well by the CART (89% correct). Only about 50% of the Good days were correctly predicted, indicating that these days are more difficult to characterize. Eight USG days occurred, but none were correctly predicted. Thus, while the Moderate (and most frequent) category was extremely well-characterized, the other categories were not. Thus, these categories need more attention by forecasters when utilizing this tool. Overall, 65% of the days were predicted correctly.

At Columbus, most days were Good, and they were well-predicted (95% correct). The remaining 25% of the days were Moderate and were predicted correctly 67% of the time. The one USG day (42.0 $\mu g/m^3$) was predicted as Moderate. Overall, Columbus was the best-predicted of all the cities, possibly due to the abundance of Good days.

Two-thirds of the test days at Detroit were Good and were correctly predicted 75% of the time. The remaining days were Moderate and were correctly predicted about 50% of the time. The one USG day (42.0 $\mu g/m^3$) was correctly predicted. Days that have the potential to be Moderate should be further evaluated by forecasters beyond the use of this tool, although Good days are well-characterized.

At St. Louis, 67% of the days were predicted correctly, which is much higher than the r^2 value of 0.56 for the CART. Moderate days were somewhat better characterized (about three-fourths of the time) compared to Good days (about two-thirds of the time). There were more Moderate days than Good, consistent with climatology, so the correct characterization of Moderate days makes sense.

At Milwaukee, more than half the days were Good and were correctly predicted 75% of the time. Moderate days were somewhat well-predicted, at 61%. Only two of the six USG days were correctly predicted. These high $PM_{2.5}$ days are infrequent; thus, they are not expected to be captured by CART. Forecaster knowledge is needed on these rare days to assess whether to predict USG.

Table 1-2. CART results for Chicago on 2003 test data. Boldface indicates the correct AQI.

Actual AQI	N	Predicted AQI	N	%	mean
Good	23	**Good**	19	83	13.0
		Mod	4	17	14.0
Moderate	55	**Mod**	43	78	22.8
		Good	12	22	20.8
Total	78		62 correct	79% correct	

Table 1-3. CART results for Cleveland on 2003 test data. Boldface indicates the correct AQI.

Actual AQI	N	Predicted AQI	N	%	mean
Good	24	**Good**	10	42	9.1
		Mod	14	58	12.7
Moderate	46	**Mod**	41	89	23.5
		Good	4	9	18.8
		USG	1	2	36.0
USG	8	**USG**	0	0	-
		Mod	8	100	47.4
Total	78		51 correct	65% correct	

Table 1-4. CART results for Columbus on 2003 test data. Boldface indicates the correct AQI.

Actual AQI	N	Predicted AQI	N	%	mean
Good	66	**Good**	62	94	9.4
		Mod	4	6	12.0
Moderate	21	**Mod**	14	67	22.6
		Good	7	33	19.1
USG	1	**USG**	0	0	-
		Mod	1	100	42.0
Total	88		76 correct	86% correct	

Table 1-5. CART results for Detroit on 2003 test data. Boldface indicates the correct AQI.

Actual AQI	N	Predicted AQI	N	%	mean
Good	60	**Good**	45	75	10.3
		Mod	15	25	11.5
Moderate	35	**Mod**	17	49	27.1
		Good	18	51	19.9
USG	1	**USG**	1	100	42.0
Total	96		63 correct	66% correct	

Table 1-6. CART results for St. Louis on 2003 test data. Boldface indicates the correct AQI.

Actual AQI	N	Predicted AQI	N	%	mean
Good	38	**Good**	23	61	11.3
		Mod	15	39	12.5
Moderate	51	**Mod**	37	73	22.9
		Good	14	27	19.9
Total	89		60 correct	67% correct	

Table 1-7. CART results for Milwaukee on a subset of 1999-2002 data. Boldface indicates the correct AQI.

Actual AQI	N	Predicted AQI	N predicted	%	mean
Good	44	**Good**	33	75	9.4
		Mod	11	25	10.9
Moderate	31	**Mod**	19	61	23.4
		Good	12	39	21.2
USG	6	**USG**	2	33	45.0
		Mod	4	67	48.4
Total	81		54 correct	67% correct	

ATTACHMENT 2: VARIABLE DEFINITIONS

Table 2-1. Surface parameter types.

Parameter Category	Parameter Abbreviation	Parameter Details
Temperature	12ZSfc_Temp	Morning temperature
	00ZSfc_Temp	Afternoon temperature
	MinOfTempC	Minimum temperature
	MaxOfTemp	Maximum temperature
	AvgDPTemp	Average daytime dew point temperature
	12ZSfc_RH	Morning relative humidity
	00ZSfc_RH	Afternoon relative humidity
Clouds	SumOfClouds	Sum of daytime cloud cover
	Clouds_day_avg	Average daytime cloud cover
Precipitation	SumOfPrecip	Sum of twenty-four hour precipitation
Pressure	12ZSfc_Pressure	Morning surface pressure
	00ZSfc_Pressure	Afternoon surface pressure
	DeltaP	Daytime change in surface pressure
Wind speed and direction	12a6aWS(WD)	Early morning average wind speed and direction
	6a12pWS(WD)	Late morning average wind speed and direction
	12p6pWS(WD)	Afternoon average wind speed and direction
	6p11pWS(WD)	Evening average wind speed and direction
	6a6pWS(WD)	Daytime average wind speed and direction
$PM_{2.5}$	prev2_xxxpm	Previous day's $PM_{2.5}$ concentration at site (i.e., MIL for Milwaukee)

Table 2-2. Upper-level parameter types.

Upper-level Parameter Category	Levels (xxx)	Parameter Abbreviation	Parameter Details
Temperature	925, 850, 700, 500	12Zxxxmb_Temp	Morning temperature aloft
		00Zxxxmb_Temp	Evening temperature aloft
		12Zxxxmb_DPTemp	Morning dew point temperature aloft
		00Zxxxmb_DPTemp	Evening dew point temperature aloft
		12Zxxxmb_RH	Morning relative humidity aloft
		00Zxxxmb_RH	Evening relative humidity aloft
	925:sfc	Tdiff925sfc12	Morning temperature difference between 925 mb and the surface
		Tdiff925sfc00	Evening temperature difference between 925 mb and the surface
	850:sfc	Tdiff850sfc12	Morning temperature difference between 850 mb and the surface
		Tdiff850sfc00	Evening temperature difference between 850 mb and the surface
Height	925, 850, 700, 500	12Zxxxmb_Height	Morning height of the pressure surface
		00Zxxxmb_Height	Evening height of the pressure surface
	925	Day_925diff	Height change at 925 mb from morning to evening
		925htdiff	24-hr morning height change at 925 mb
	500	500htdiff	24-hr morning height change at 500 mb
Wind	925, 850, 700, 500	12Zxxxmb_WS	Morning aloft wind speed
		12Zxxxmb_WD	Morning aloft wind direction
		00Zxxxmb_WS	Evening aloft wind speed
		12Zxxxmb_WD	Evening aloft wind direction

ATTACHMENT 3: REFERENCES

Brown S.G. and Dye T.S. (2003) $PM_{2.5}$ forecasting – synoptic typing. Technical memorandum prepared for Lake Michigan Air Directors Consortium, Des Plaines, IL, by Sonoma Technology, Inc., Petaluma, CA, STI 903480-2414-TM, September.

MacDonald C.P., Dye T.S., Strohm D.E., Nguyen D.T., Miller D.S., and Ryan W. (2003) $PM_{2.5}$ regional forecasting workshop (San Francisco, Research Triangle Park, Chicago, Boston). Workshop prepared for U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, Research Triangle Park, NC, by Sonoma Technology, Inc., Petaluma, CA, STI-902466-2411, September.

National Research Council (1991) *Rethinking the Ozone Problem in Urban and Regional Air Pollution*, National Academy of Sciences/National Research Council, National Academy Press, Washington, DC.

Seinfeld J.H. and Pandis S.N. (1998) *Atmospheric chemistry and physics: from air pollution to global change*, J. Wiley and Sons, Inc., New York, New York.

Stoeckenius T. (1990) Adjustment of ozone trends for meteorological variation. Presented at the *Air and Waste Management Association's Specialty Conference, Tropospheric Ozone and the Environment, Los Angeles, CA, March 19-22*.

U.S. Environmental Protection Agency (2003) Guidelines for developing an air quality (ozone and $PM_{2.5}$) forecasting program. U.S. Environmental Protection Agency, Research Triangle Park, NC. Prepared by Dye T.S., MacDonald C.P., Anderson C.B., Hafner H.R., Wheeler N.J.M., and Chan A.C. at Sonoma Technology, Inc., Petaluma, CA, 902461-2295-DFR2, January.