Application of Prognostic Meteorological Variables to Forecasts of Daily Maximum One-Hour Ozone Concentrations in the Northeastern United States

TERRY L. CLARK[1] AND THOMAS R. KARL[2]

Environmental Sciences Research Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, NC 27711

(Manuscript received 20 June 1980, in final form 12 July 1982)

ABSTRACT

A linear multiple regression equation was developed for each of 27 ozone monitoring sites in the northeastern United States to forecast the next day's maximum 1 h average ozone concentration. Thirty-five prognostic meteorological variables, the climatological daily maximum surface temperature, the length and direction of 12 and 24 h backward trajectories, and three air quality variables relating to the seasonality or the upwind ozone concentrations were considered as possible predictors in each of the regression equations. Data pertaining to 244 randomly selected days formed the developmental or the dependent data set, while the data pertaining to the remaining 122 days in the months of June, July, August and September of 1975, 1976 and 1977 were used to assess the performance of the regression equations. Performance was assessed and compared to that of persistence, *via* statistical evaluations of site-specific forecasts. In addition, areas of the Northeast where the 1 h ozone standard was predicted to be exceeded, were compared to the areas where the standard was exceeded.

The results indicated that approximately half of the predictions generated from the independent data set were within 20% of the observations, while 77% were within 40% of the observations. A tendency for the underprediction of the maximum concentrations was noted. Overall, the regression equations performed best in forecasting the trends and patterns of the daily 1 h average ozone concentrations.

1. Introduction

In the absence of errors realized from the measurement technique, observations are a direct result of the values of pertinent physical parameters, each influencing the observation in varying degrees. For instance, the daily maximum surface temperature is dependent on a set of parameters, which includes the amount of cloud cover and the boundary-layer temperature, wind velocity and relative humidity. One parameter, say boundary-layer temperature, would exert more influence on the maximum surface temperature than other parameters.

Multiple regression analysis attempts to identify those parameters influencing the observation, and to weight each parameter according to its "typical" influence on the observation. Unfortunately, the analysis only considers those parameters, from a list of parameters for which there are data available. If, for some reason, a pertinent parameter was omitted from the list, the success of the analysis would suffer, especially if the parameter was wholly independent of the others.

The advantages of multiple regression analysis techniques over dynamic methods include its objectivity, its usefulness in forecasting fields of numerical variables, and its ability to identify the pertinent parameters. Moreover, this analysis technique does not require the severe physical restrictions that the dynamic methods require (Panofsky and Brier, 1965).

Several decades ago, Grant (1956) studied the use of multiple regression analysis to forecast the values of meteorological variables. Since then, multiple regression analysis techniques have been applied by Klein (1963) and Paegle (1973) to develop precipitation forecast tools based on the upper-air flow patterns; by Harnack (1979) to forecast winter temperatures in the central and eastern United States; by Glahn and Lowry (1972) to forecast surface winds, probability of precipitation, maximum surface temperature, cloud amount, and conditional probability of frozen precipitation; by Meisner (1979) to temporally extrapolate rainfall amounts; by Guldberg *et al.* (1977) to assess air quality impact from an altered emission scenario; and by Karl (1979) to forecast, 24 and 48 h in advance, the daily maximum 1 h average ozone concentrations at three different groups of monitoring sites in metropolitan St. Louis, MO.

During the study described in this paper, a multiple regression analysis technique was applied using available National Weather Service (NWS) forecast data and Environmental Protection Agency (EPA) air

[1] On assignment from National Oceanic and Atmospheric Administration, U.S. Department of Commerce.

[2] Currently affiliated with the National Climatic Center, National Oceanic and Atmospheric Administration, U.S. Department of Commerce.

quality data to develop linear multiple regression equations to forecast the daily maximum 1 h average surface ozone concentrations during the summer months at 27 monitoring sites in the northeastern United States. The application of the technique and the performance of the regression equations (using an independent data set) are discussed.

This study is related to the studies of Karl (1979) and Wolff and Lioy (1978) in that linear multiple regression equations were developed to relate the maximum 1 h average ozone concentrations to meteorological variables. However, this study differs from that of Karl, since in the former a regression equation was developed for each monitoring site. This study differs from that of Wolff and Lioy, since they developed only one regression equation from the values of a few meteorological parameters 24 h upwind of New Jersey to forecast the average maximum ozone concentration at four widely varying sites in that state.

2. Data

The data set used in the regression analysis included daily air quality data, climatological data, prognostic meteorological data and trajectory analysis data. The tables in this section list the variables and the validation time for the prognostic variables. Each variable was assigned a code number which was used in a subsequent section.

The data were either sampled or calculated at 27 sites in the northeastern United States during the days of June, July, August and September of 1975, 1976 and 1977. Two-thirds of the days (244 days) were randomly selected for the developmental sample, while the remaining one-third (122 days) formed the test or independent sample.

a. Ozone data.

Daily 1 h ozone concentrations measured at 27 surface monitoring sites were obtained from the

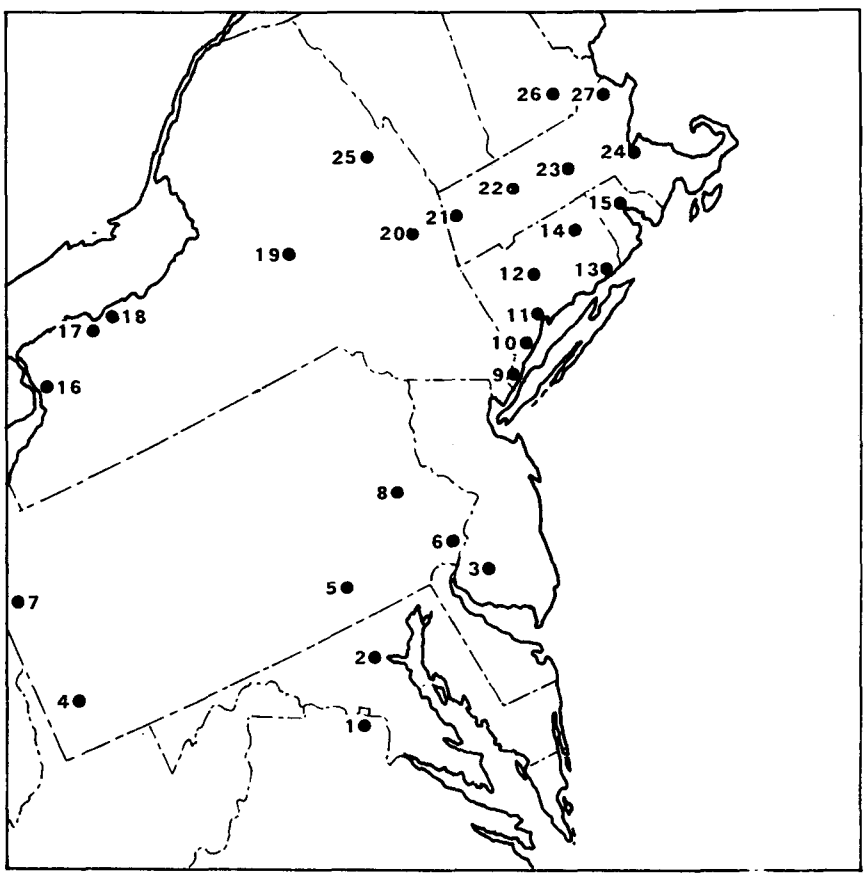

FIG. 1. Location of the 27 SAROAD monitoring sites used in this study. Sites: 1, Seven Corners, VA; 2, Baltimore, MD; 3, Ancora, NJ; 4, Charleroi, PA; 5, York, PA; 6, Philadelphia, PA; 7, New Castle, PA; 8, Bethlehem, PA; 9, Greenwich, CT; 10, Bridgeport, CT; 11, New Haven, CT; 12, Hamden, CT; 13, Groton, CT; 14, Eastford, CT; 15, Providence, RI; 16, Buffalo, NY; 17, Rochester, NY; 18, Monroe County, NY; 19, Utica, NY; 20, Rensselaer, NY; 21, Pittsfield, MA; 22, Amherst, MA; 23, Worcester, MA; 24, Quincy, MA; 25, Glens Falls, NY; 26, Manchester, NH; 27, West Newbury, MA. Sites 3, 9, 14, 18, 22 and 27 are rural or remote.

TABLE 1. Prognostic LFM variables available in the regression analysis for each site. Validation times of the variables are indicated by the asterisks in the appropriate columns. The boundary layer is the lowest 50-mb layer; layer 1 is the layer between the lowest 50 mb and 720 mb; and layer 2 is the layer between 720 and 490 mb. The u and v components of the wind are based on the LFM coordinate system. RH is relative humidity and B. L. boundary layer.

Variable code	Prognostic variable	Validation time after 0000GMT 0000 a	0600 b	1200 c	1800 d	2400 e	Units
4	700-mb height	*		*	*	*	m
5	850-mb temperature			*			K
6	700-mb temperature			*		*	K
7	B. L. u-component		*	*	*	*	m s^{-1}
8	B. L. v-component		*	*	*	*	m s^{-1}
9	B. L. wind speed		*	*	*	*	m s^{-1}
10	850-mb u-component		*	*	*	*	m s^{-1}
11	850-mb v-component		*	*	*	*	m s^{-1}
12	850-mb w-component			*		*	μb s^{-1}
13	850-mb wind speed		*	*	*	*	m s^{-1}
14	700-mb u-component			*		*	m s^{-1}
15	700-mb v-component			*		*	m s^{-1}
16	700-mb w-component			*	*	*	μb s^{-1}
17	700-mb wind speed			*		*	m s^{-1}
18	Sea-level pressure	*		*		*	mb
19	B. L. RH	*		*	*	*	%
20	Layer 1 RH		*	*	*	*	%
21	Layer 2 RH		*	*	*	*	%
22	Entire layer RH		*	*	*	*	%

EPA Storage and Retrieval of Aerometric Data (SAROAD) System. Fig. 1 illustrates the location of the sites where maximum ozone concentrations were measured and predicted in this study. The sites were selected on the basis of the availability of continuously monitored summertime ozone data. Only six of these sites were classified as being rural or remote, while the remainder were classified as being urban or suburban. Many of the sites where ozone was continuously monitored during 1975–77 were either urban or suburban.

The square root of the maximum 1 h average ozone concentration was used as the predictand in the regression analysis. This more closely satisfied the condition of normally distributed residuals, a condition imposed by the F-test used in this study.

Three independent variables in the regression analysis were based on the ozone data itself. One variable was the average daily maximum ozone concentration (variable code 1), which reflected the seasonality of maximum ozone concentrations. This variable was computed for each site by first calculating the average maximum ozone concentration for every five calendar-day period in the summer months of 1975–77. Each five calendar-day average considered ozone data from five days of each year or a total of fifteen days. Each average value was assumed to represent the average maximum ozone concentration for the third day of each five calendar-day period. The average maximum ozone concentrations for the remaining calendar days were determined by linear interpolation or extrapolation, whichever was appropriate.

The other two variables were the estimated maximum ozone concentration (variable code 2) and the category of the maximum ozone concentration (variable code 3) at the origin of 24 h trajectories ending at the 27 SAROAD sites. These trajectories were determined *via* a technique developed by Heffter and Taylor (1975) using NWS rawinsonde data obtained twice daily and averaged across the 900–1100 m layer. Winds in this layer were chosen to represent the transport wind, since the layer is in the middle of the boundary layer. The maximum ozone concentration at a trajectory's point of origin was estimated from the ozone data measured at ten SAROAD sites *via* an inverse distance squared weighting scheme. These ten sites were Babylon, Amherst and Glens Falls, NY; Seven Corners, VA; Middletown, OH; De Soto County, MS; Rock Hill, SC; Jackson, MI; Milwaukee, WI; and Des Moines, IA. Admittedly, these values were crudely determined, but they should be correlated with the ozone levels transported into the northeastern United States.

b. Meteorological data.

A set of climatological and prognostic meteorological data was compiled for each of the 27 SAROAD ozone monitoring sites. Each set of data was used in the analysis to 1) determine what group of variables was most related to the maximum ozone concentration, and 2) develop regression equations to forecast the maximum ozone concentrations. The list of variables was compiled on the basis of the knowledge of relationships between meteorological variables and

ozone concentrations, the availability of the data, and speculation.

Forty meteorological variables were available as predictors in the regression analysis. Most of the prognostic variables were computed by the NWS Limited-Area Fine Mesh (LFM) Model, which is operationally executed twice each day. They included pressure heights, temperatures, wind speeds and relative humidities in the lower troposphere. These variables listed in Table 1, were obtained from the model runs, using raw data collected near sunset the day prior to the forecast day. A linear spatial interpolation scheme was used to determine the value of the variables at each of the 27 sites.

A few variables, namely, the prognostic maximum and minimum surface temperatures and the probability of precipitation, were obtained from Model Output Statistics (see Glahn and Lowry, 1972). These variables, listed in Table 2, were determined by the NWS *via* an objective forecasting technique based on statistical relationships between a predictand and prognostic meteorological variables from the LFM model.

TABLE 2. Prognostic Model Output Statistics (MOS) available in the regression analysis for each site.

Variable code	Prognostic variable	Validation time after 0000 GMT	Units
23	Maximum surface temperature	1200–2400	°F
24	Minimum surface temperature	2400–3600	°F
25	Probability of precipitation	1200–2400	%
26	Probability of precipitation	2400–3600	%

Additional meteorological variables available in the analysis are listed in Table 3. The "quadrant of backward trajectory approach" parameterized the direction from which ozone was transported to the monitoring site. The 0–90° quadrant was referred to as the northeast quadrant, 90–180° as the southeast quadrant, and so on. A variable associated with each of the quadrants was assigned one of three values, depending on the location of the 12 and 24 h backward trajectories. If both the 12 and 24 h backward trajectories were located in only one quadrant, the variable associated with that quadrant was assigned a value of 1.0. If either backward trajectory was located in more than one quadrant at any time, the value associated with each of those quadrants was assigned a value of 0.5. A value of 0.0 was assigned to the variable associated with the quadrant containing neither backward trajectory.

3. Development of the forecasting technique

a. Multiple linear regression analysis.

A computer program available from the Health Sciences Computing Facility of the School of Medicine, University of California, Los Angeles (Dixon, 1977), incorporating a backward and forward stepwise regression technique, was used to develop daily maximum ozone forecast equations for each of the 27 monitoring sites. Essentially, the technique computed a multiple linear regression equation in a stepwise fashion. At each step, the predictand y was estimated by

$$y' = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_i ,$$

where α_0 is the intercept, α_i are the coefficients of the predictors X_i, and p is the number of predictors in the equation. The coefficients were determined by the least squares method.

At each step, a variable was either added (forward stepping) or subtracted (backward stepping) and replaced by another in the equation. Selection and removal of variables as predictors were based on their F values computed at each step. The variable with

TABLE 3. Remaining meteorological variables available in the regression analysis for each site.

Variable code	Meteorological variable. Validation time after 0000 GMT shown in parentheses.	Units
27	Average daily maximum surface temperature	°F
28	Length of 24 h backward trajectory	km
29	Quadrant of backward trajectory: NE	—
30	Quadrant of backward trajectory: SE	—
31	Quadrant of backward trajectory: SW	—
32	Quadrant of backward trajectory: NW	—
33	Square root of deviation of maximum surface temperature from normal	°F$^{1/2}$
34	Square root of deviation of minimum surface temperature from normal	°F$^{1/2}$
35	Entire layer relative humidity (2400 minus 0600)	%
36	Sea-level pressure (2400 minus 0000)	mb
37	850-mb temperature (1200) minus maximum surface temperature (1200 − 2400)	K
38	700-mb temperature (1200) minus maximum surface temperature (1200 − 2400)	K
39	850- and 700-mb average temperature minus maximum surface temperature (1200 − 2400)	K
40	700-mb temperature (2400) minus maximum surface temperature (1200 − 2400)	K
41	Boundary layer relative humidity (2400 minus 0600)	%
42	Boundary layer wind speed (2400 minus 0600)	m s^{-1}
43	850-mb wind speed (2400 minus 0600)	m s^{-1}
44	Layer 2 relative humidity (2400 minus 0000)	%
45	850 and 700-mb average temperature	K
46	700-mb height (2400 minus 0000)	m
47	700-mb w-component (2400 minus 0600)	μb s^{-1}

the highest F-to-enter value was selected as a predictor during each step. This value is defined as

$$F\text{-to-enter} = \frac{\text{SS(current)} - \text{SS(next)}}{\text{SS(next)}/(n - p - 1)},$$

where SS(current) is the sum of squares of the residuals (observation − prediction) at the current time step, SS(next) is the sum of squares of the residuals after the variable is added to the equation, n is the sample size, and p is the number of variables considered in the study.

A predictor can be removed from the regression equation at any step if the F-to-remove value was less than a user-specified value. The F-to-remove value is defined as

$$F\text{-to-remove} = \frac{\text{SS(removed)} - \text{SS(not removed)}}{\text{SS(not removed)}/(n - p)}.$$

If no predictor was removed from the regression equation during a step, a predictor was exchanged with a variable not yet in the equation, if the exchange increased the multiple linear correlation between the predictors and the predictand. The stepping procedure was ended when the F-to-enter values of the variables not selected as parameters in the regression equation were less than a user-specified value.

b. Predictand enhancement.

There is a tendency for regression forecast techniques (and most other techniques) to predict values over a smaller range than the range of observed values. These techniques tend to overestimate the lowest observed values and underestimate the highest observed values. One reason for this tendency is that the extreme observed values occur much less often than the values approximating the averaged observed value. However, accurately forecasting the extreme values is usually of the utmost concern, especially regarding daily maximum 1 h average ozone concentrations.

In the literature, there have been several references to algorithms, which when applied to forecasts, are capable of enhancing the accuracy of the forecast technique by reducing the overestimation of the lower and the underestimation of the higher observed values (Klein *et al.*, 1962; Russo *et al.*, 1964; Glahn and Allen, 1966; Bennett and Leslie, 1979). A similar form of the prediction enhancement equation used by Glahn and Allen,

$$P_E = \frac{P - \bar{P}}{R} + \bar{P},$$

was applied to each of the 27 regression equations developed in this study. In the equation, P_E is the enhanced predictand, P the predictand obtained from the regression equation, $\bar{P}$ the average value of the predictands determined from the dependent data sample, and R the multiple correlation coefficient of the predictand with the predictors in the particular regression equation.

Essentially, the more the predictand value deviated from the average predictand value, the greater was the enhancement in most cases. To illustrate the prediction enhancement concept, suppose the $\bar{P}$ and R calculated from the dependent data set for one of the sites were 126 $\mu g\ m^{-3}$ and 0.61, respectively. Now suppose that the regression equation predicted 200 $\mu g\ m^{-3}$ to be the next day's maximum 1 h average ozone concentration. After the prediction enhancement equation for that site had been applied, the new prediction would be 247 $\mu g\ m^{-3}$. Chances are good that this new prediction would be closer to the observed value.

4. Results and evaluations

The maximum 1 h average ozone forecast equations developed for each of the 27 sites are presented in Table 4. Each of these regression equations consisted of an intercept and from 5 to 14 predictor terms. The predictor terms are listed in the order of their impact on the magnitude of the predictand. An asterisk appears behind one of the prediction terms signifying that the intercept and the prediction terms preceeding the asterisk explain 90% or more of the value of the predictand.

The predictor terms relating to atmospheric temperature or pressure had the greatest impact on the predictand in all but two equations. In these two equations, the average daily maximum ozone concentration had the greatest impact. Table 5 lists the variables appearing in the 27 regression equations most often. The maximum surface temperature was selected as a predictor in 19 of the equations. All of the coefficients associated with this parameter were >0, indicating that, in general, the greater the maximum surface temperature, the higher was the maximum 1 h average ozone concentration.

The "quadrant of backward trajectory approach" variables were selected in nine different regression equations for a total of 12 times. However, the fact that they were selected so often should not be interpreted as a measure of their significance as predictors. These parameters appeared near the end of the list of prediction terms in each equation, reflecting their relative unimportance. Only two of these 12 prediction terms appeared before the asterisks in Table 5.

The prognostic sea-level pressure valid at 0000 GMT the next day and the length of the 24 h backward trajectory were both selected nine times. The coefficients associated with both parameters were <0, implying that the greater the sea-level pressure, the lower the maximum 1 h average ozone concentration and the lower the transport wind speed, the greater the maximum 1 h average ozone concentration. It should be noted that, with few exceptions, parameters relating to the sea-level pressure did not

TABLE 4. The regression equations developed to predict the square root of the maximum 1 h average ozone concentrations at the 27 northeastern United States monitoring sites and the statistical evaluations of the square of the predictions of the regression equations (R.E.) for the independent data set and persistence (Pers.). Each of the predictor terms in the regression equations includes a variable code in parentheses; these variables are defined in Section 2. The intercept and all the predictor terms preceding the asterisk explain 90% of the predicted maximum concentration.

Site number	Average observed maximum ($\mu g\ m^{-3}$)	Average predicted maximum ($\mu g\ m^{-3}$)	Multiple correlation coefficient		rms error		Contingency table χ^2		Intercept	Regression equation terms			
			R.E.	Pers.	R.E.	Pers.	R.E.	Pers.					
1	156	158	0.71	0.48	50	62	57	11	2.168	+ 0.134 (23)	− 0.120 (19a)	+ 0.091 (19c)	+ 0.022 (1)
										− 0.027 (25)	− 0.167 (9e)	− 0.134 (17c)	+ 0.117 (7b)*
										− 0.026 (35)	− 0.035 (41)	− 0.003 (28)	− 0.336 (16d)
2	145	138	0.54	0.51	81	77	18	29	−141.429	+ 0.129 (18a)	+ 0.286 (23)*	+ 0.278 (7d)	− 0.409 (13d)
										− 0.128 (11b)	+ 0.141 (8e)		
3	147	129	0.56	0.37	45	60	22	8	27.627	− 0.033 (4a)	+ 0.024 (4c)	+ 0.239 (23)	+ 0.135 (40)*
										− 0.037 (19c)	+ 0.168 (10e)	− 0.154 (17c)	− 0.169 (13e)
										+ 1.54 (29)	− 0.074 (36)		
4	154	143	0.70	0.73	52	51	26	30	−30.968	− 0.048 (1)	+ 0.697 (27)	+ 0.419 (33)	− 0.035 (22e)
										− 0.155 (9b)	− 1.999 (30)*	+ 0.063 (14c)	− 0.004 (28)
5	175	182	0.69	0.48	60	69	26	30	−2.753	+ 0.167 (23)	+ 0.109 (24)	+ 0.099 (19c)	− 0.082 (20c)
										+ 0.087 (20d)	− 0.080 (22e)	− 0.007 (1)	+ 2.063 (29)
										− 1.548 (30)	− 0.088 (17c)*	− 0.004 (28)	+ 0.698 (32)
6	140	122	0.53	0.26	80	90	21	9	3.236	+ 0.033 (1)	− 0.519 (38)	+ 0.460 (37)	− 0.487 (13e)
										− 0.064 (18)	− 0.373 (17c)*	+ 0.345 (17e)	+ 0.169 (7d)
										+ 0.229 (8d)			
7	112	112	0.60	0.31	49	55	19	5	−8.968	+ 0.468 (5c)	− 0.384 (6c)	+ 0.176 (27)	− 0.044 (22e)
										− 0.225 (13d)*	− 0.023 (19a)	+ 0.083 (8e)	
8	202	174	0.67	0.51	83	90	27	22	−94.180	+ 0.168 (18a)	− 0.088 (18e)	+ 0.282 (23)*	+ 0.031 (1)
										− 0.061 (22c)	− 0.200 (13d)		
9	184	156	0.74	0.53	85	117	28	23	−3.797	+ 0.729 (23)	− 0.454 (45)	+ 0.632 (37)	− 0.064 (19a)
										+ 0.356 (8d)	− 0.375 (13c)	+ 0.165 (10b)	+ 0.300 (10c)
										− 0.229 (10e)*	− 0.059 (35)	+ 2.484 (29)	
10	176	151	0.72	0.42	78	108	50	9	−6.728	+ 0.258 (23)	− 0.275 (13d)	− 0.160 (11c)	+ 0.217 (8d)*
										+ 0.166 (10d)			
11	166	134	0.68	0.73	90	120	36	30	−8.388	+ 0.265 (23)	+ 0.337 (10b)	+ 0.187 (8e)	− 0.166 (13b)
										− 0.186 (17e)	− 0.133 (36)*	+ 1.624 (29)	
12	167	145	0.61	0.41	80	95	31	22	597.325	− 0.857 (18e)	+ 0.097 (4e)*	+ 0.149 (24)	− 0.481 (6e)
										+ 0.025 (11b)	− 0.078 (22e)	− 0.153 (13e)	+ 0.051 (41)
13	178	177	0.62	0.42	98	95	32	11	2.461	+ 0.593 (5c)	+ 0.347 (33)	− 0.383 (6c)	− 0.083 (19c)
										+ 0.278 (10c)	− 0.255 (13d)	+ 0.191 (7b)	+ 0.162 (8e)
										− 0.041 (35)*	− 0.877 (12e)	− 0.121 (9e)	− 0.004 (28)
14	155	135	0.66	0.50	67	81	32	25	206.062	− 0.255 (18e)	+ 0.022 (4d)*	− 0.055 (19d)	+ 0.035 (33)
										+ 0.070 (8b)	+ 0.175 (10b)	+ 0.089 (42)	
15	132	127	0.70	0.45	59	78	17	18	111.158	− 0.112 (18e)	+ 0.337 (23)	− 0.128 (27)*	− 0.078 (19e)
										+ 0.031 (20d)	− 0.151 (10e)	+ 0.143 (14c)	− 0.003 (28)
										− 0.630 (16e)			
16	134	181	0.55	0.46	65	57	26	20	4.737	+ 0.173 (23)	− 0.048 (22e)	+ 0.114 (37)	− 0.027 (19b)
										− 0.213 (17c)	+ 0.154 (17e)*	+ 0.018 (26)	+ 0.111 (7b)
17	119	113	0.68	0.51	40	49	48	25	0.796	+ 0.152 (23)	− 0.040 (20c)	+ 0.133 (7b)	+ 0.130 (8b)
										+ 0.018 (21a)	− 0.059 (14c)*	+ 0.149 (43)	
18	124	114	0.66	0.49	45	49	33	12	7.592	− 0.192 (27)	+ 0.016 (21a)	+ 0.119 (23)	+ 0.093 (6e)
										− 0.052 (19c)	+ 0.093 (6e)*	+ 0.132 (7c)	+ 0.082 (8c)
										+ 0.123 (43)			
19	105	102	0.65	0.45	36	43	42	19	8.400	+ 0.144 (23)	− 0.083 (27)	+ 0.212 (7b)	− 0.026 (19d)
										− 0.182 (13b)	+ 0.100 (11b)	+ 0.076 (8c)*	− 0.018 (21e)
										− 0.074 (36)			
20	120	126	0.61	0.39	55	54	43	9	175.780	− 0.234 (18e)*	− 0.114 (38)	+ 0.112 (7b)	+ 0.109 (8b)
										− 0.022 (19a)	+ 0.414 (16c)	− 0.426 (16e)	− 0.028 (20e)
										− 0.004 (28)	+ 0.685 (31)		
21	113	107	0.66	0.39	41	56	32	10	50.782	− 0.051 (18e)	+ 0.173 (23)*	− 0.039 (19d)	+ 0.094 (11b)
										+ 0.112 (42)	+ 1.067 (31)	− 0.509 (12e)	
22	123	107	0.60	0.43	61	67	36	17	3.250	+ 0.151 (23)	− 0.069 (19a)	− 0.315 (11d)	+ 0.253 (11e)
										+ 0.163 (7b)	+ 0.039 (19c)	+ 0.192 (8d)	− 0.024 (26)
										+0.117 (11b)	− 1.070 (12e)	− 0.135 (17e)	− 0.041 (35)*
										− 0.003 (28)	+ 0.022 (44)		
23	120	112	0.72	0.59	44	61	47	20	84.716	− 0.082 (18e)	+ 0.189 (23)*	− 0.047 (19a)	+ 0.123 (8b)
										− 0.027 (20e)	− 0.218 (13c)	+ 0.091 (13c)	+ 0.707 (31)
24	134	135	0.70	0.45	59	68	44	12	304.794	− 0.420 (18e)	+ 0.042 (4e)	+ 0.271 (23)*	− 0.219 (6e)
										+ 0.149 (37)	− 0.058 (19d)	− 0.024 (1)	+ 0.109 (7b)
										− 0.136 (13b)			
25	93	77	0.59	0.53	47	44	31	29	81.762	− 0.068 (18e)	+ 0.519 (33)*	+ 0.119 (8b)	− 0.040 (19d)
										− 0.137 (9d)	− 0.003 (28)	+ 1.318 (31)	+ 1.10 (32)
26	104	97	0.68	0.49	40	53	32	23	37.954	− 0.011 (4d)	+ 0.136 (23)	− 0.033 (19a)	− 0.205 (17c)*
										− 0.145 (7c)	− 0.018 (21c)	− 0.092 (9b)	− 1.472 (30)
27	135	137	0.47	0.40	90	66	15	13	39.224	− 0.012 (4e)	+ 0.306 (23)	− 0.047 (1)	− 0.078 (22c)*
										− 0.202 (13b)	− 0.103 (14c)	− 0.003 (28)	

TABLE 5. The variables most often selected as predictors in the 27 regression equations.

Variable code	Variable	Validation time after 0000GMT	Number of times selected
23	Maximum surface temperature	1200–2400	19
29–32	Quadrant of backward trajectory	0000–2400	12
18e	Sea-level pressure	2400	9
28	Length of 24 h trajectory	0000–2400	9
7b	Boundary layer *u*-component	0600	7
1	Average daily maximum ozone	—	7
19a	Boundary layer relative humidity	0000	7

appear in the regression equations containing the maximum surface temperature parameter.

In eight of the regression equations, the prognostic boundary layer u wind component valid at 0600 GMT appeared. The coefficients associated with this parameter were >0, implying that the maximum 1 h average ozone concentrations generally increased as the eastward wind component increased (a west wind versus a north wind). This also implied that, in general, the maximum 1 h average ozone concentrations were higher when air masses were transported across the region from the Midwest, than they were when the air masses were transported from eastern Canada.

It is interesting to note that the variables relating to the maximum 1 h average ozone concentrations upwind of the 27 sites on the previous day were selected as predictors in only four equations (for sites 7, 16, 20 and 26). In each of these equations, these variables were the last ones selected prior to the completion of each equation. Since the order of selection was directly related to the significance of the variables as predictors, these variables were insignificant. The effort of computing the values of these variables for future applications would not be worthwhile; therefore, they were omitted from the regression equations in Table 4.

The regression equations were applied to the independent data set to assess the forecasting utility. No forecasts were made at sites for the days when essential data were unavailable. Several statistical parameters were calculated from this application and are presented in Table 4. In addition, statistical parameters were calculated for a persistence forecast (i.e., the maximum 1 h average ozone concentration measured the preceding day).

The multiple correlation coefficient and the rms error provide a means of comparing observational values to the forecast values. Although the average multiple correlation between the observations and the forecasts was only 0.64 for the regression equations, that for the persistence forecast was much lower (0.47). Only at New Haven and Charleroi was the multiple correlation greater for the persistence forecast (0.73 versus 0.68 and 0.73 versus 0.70, respectively). The rms error was significantly greater (at least by 20) for the regression equation forecasts at only two sites—Buffalo (106 versus 57 $\mu g\ m^{-3}$) and West Newbury (90 versus 66 $\mu g\ m^{-3}$).

Contingency tables provide a means of determining the ability of the forecasting procedure to forecast the relative magnitude of a variable (i.e., low, moderate and high). For each site, a pair of three-by-three contingency tables was generated, one based on the regression equations and the other based on the persistence. The range and limits of the three classes were determined so that the class size would be nearly equal. For each contingency table, the χ^2 value was calculated and compared to the critical χ^2 value at the 5% level with four degrees of freedom (9.49). If the calculated value of χ^2 was <9.49, the forecasting skill could have been attained by chance. When the calculated value of χ^2 exceeded the critical value, the degree of assurance that the forecasting technique performed better than chance, increased with increasing values of χ^2.

All of the values of χ^2 associated with the regression equation contingency tables exceeded the critical value, while five of the persistence contingency tables yielded values of χ^2 below the critical value. For all but four sites, the values of χ^2 for the regression contingency tables exceeded those for the persistence contingency tables, indicating that the skill of the persistence forecasts was closer to that of chance.

As another means of assessing the performance of the regression equations as forecasting tools, a fre-

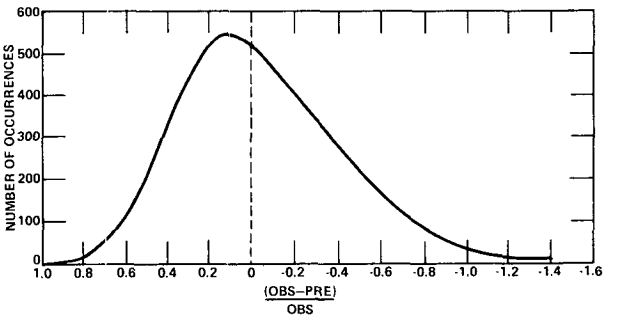

FIG. 2. Number of occurrences in each class of size 0.2 of the normalized residuals for the maximum 1 h average ozone concentrations predicted and observed at all sites using the independent data set only.

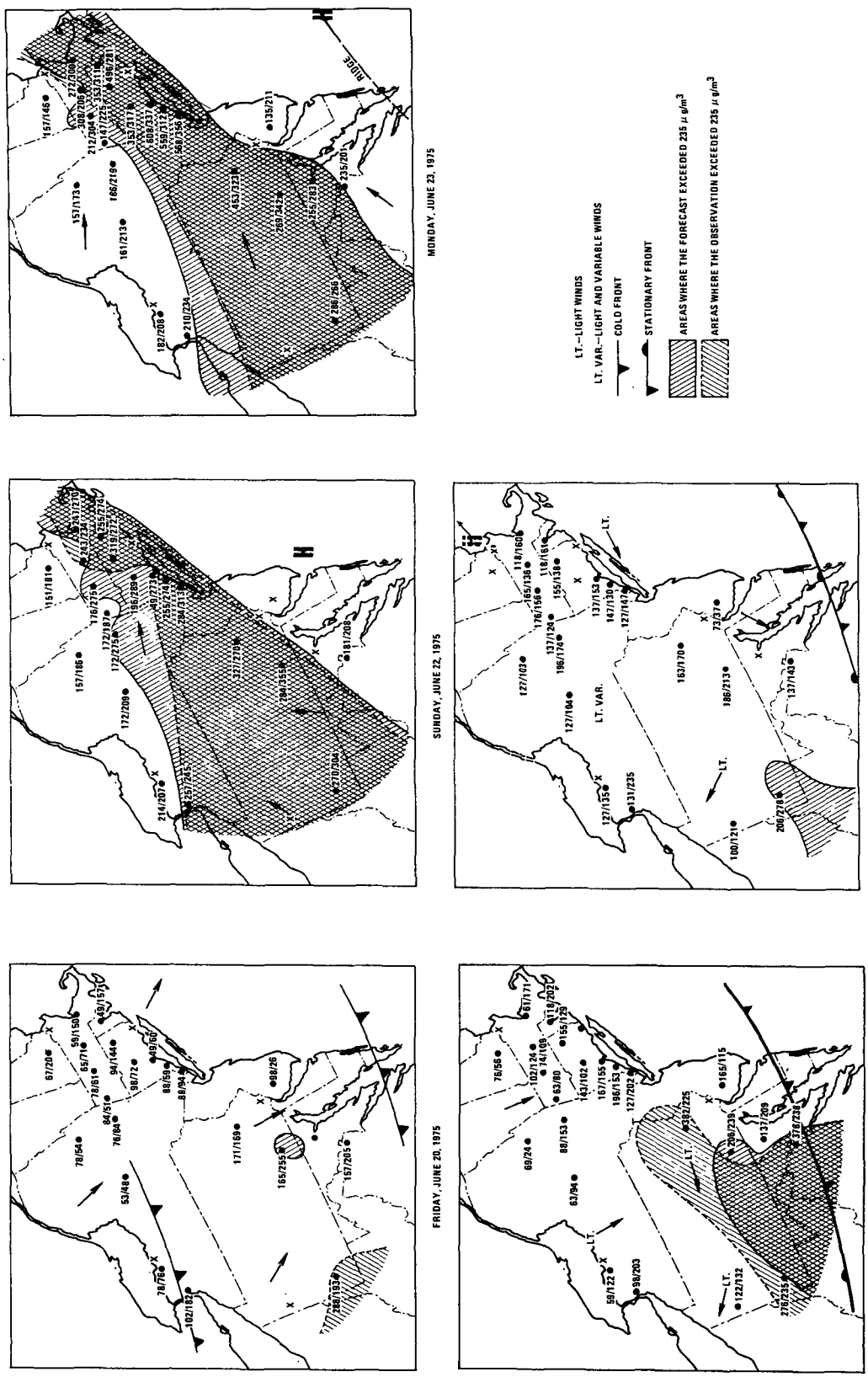

FIG. 3. Maps of observed (numerator) and predicted (denominator) maximum 1 h average ozone concentrations ($\mu g\ m^{-3}$) for each of the days in the independent data set during the 20–25 June 1975 ozone episode. The areas where the observed and predicted maximum concentrations exceeded 235 $\mu g\ m^{-3}$ are shaded. The general surface wind flow and the location of pressure centers and fronts at 0700 EST are included.

quency plot of normalized residuals was constructed using class sizes of 0.2 (Fig. 2). The predicted values at all the sites and for all the days in the independent data set were considered. The plot indicated that nearly half of the predictions were within 20% of the observations, while 77% were within 40%. Fig. 2 also illustrates the fact that the regression equations, even when the prediction enhancement technique was applied, tended to underpredict the maximum 1 h average ozone concentration.

Perhaps the most useful information any forecasting technique could provide is the likelihood the maximum 1 h average ozone concentration would exceed the 1 h standard of 235 $\mu g\ m^3$. The regression equations do not directly estimate the likelihood of an exceedance, but the magnitude of the predicted value does imply the likelihood. Quite simply, if the predicted maximum 1 h average concentration at a given site was 300 $\mu g\ m^{-3}$ for one day and 50 $\mu g\ m^{-3}$ for the next day, the likelihood of an exceedance on the first day would be greater than that for the second day.

To assess the ability of the regression equations to predict these exceedances, several series of maps of observed and predicted maximum 1 h average ozone concentrations were analyzed for days in the independent data set preceeding, during and following multiple-day episodic events. The areas where the forecasting technique indicated the 1 h standard would be exceeded and the areas where the 1 h standard was exceeded were compared. Fig. 3 presents the series of maps for the four independent days during an ozone episode that occurred 20–25 June 1975. The map for 26 June was included in this series to show the results of the regression forecasting technique the day the episode ended. Of the 366 days examined in this study, the highest ozone concentrations were measured on 23 June 1975.

During the first day of this period, two weak and rapidly moving cold fronts moved across the northeastern United States bringing northerly and northwesterly winds and moderate amounts of rainfall to the region, with the exception of New England. Despite the frontal passages and the wind directions, the maximum surface temperatures in the Washington–Boston Corridor surpassed 30°C. Only the Charleroi site, located in the southwestern corner of Pennsylvania, violated the 1 h ozone standard. The forecasting technique, however, did not predict this to occur. The forecasting technique did predict that the York site would violate the standard, although the maximum 1 h average ozone concentration there reached only 165 $\mu g\ m^{-3}$.

The next day in this period included in the independent data set was 22 June. After the center of the high pressure system moved rapidly across the region bringing much cooler temperatures to most of the Northeast the day before, the winds on this day were westerly and southwesterly. With the exception of the sites near the center of the pressure system, maximum surface temperatures across the region increased 5°C from the previous cool day. The ozone concentrations rapidly increased over much of the region. The maximum 1 h average ozone concentrations at Bethlehem and Eastford exceeded 300 $\mu g\ m^{-3}$. Ten of the 21 sites for which data were available this day, monitored ozone concentrations exceeding the 1 h standard. Nine of these 10 sites were identified by the forecasting technique to exceed the 1 h standard. The observed maximum at the Worcester site was only 8 $\mu g\ m^{-3}$ over the standard, while the prediction was only 1 $\mu g\ m^{-3}$ under the standard.

On 23 June, the maximum surface temperatures across the region increased slightly. The winds in the southern and central portions of the region remained from the west and southwest, while the winds in the remaining portion were from the northwest. The maximum concentrations again increased sharply: maximum 1 h average concentrations at the Bethlehem and western Connecticut sites exceeded 350 $\mu g\ m^{-3}$. On this day, 11 sites exceeded the standard and all but the Worcester site were identified by the forecasting technique. The Amherst site was incorrectly identified to exceed the standard. Both of these sites were very close to the border of the area correctly forecast to exceed the standard.

On 25 June, a weak cold front quickly traversed the region, creating light shower activity and northerly and easterly winds. However, maximum surface temperatures in the Washington–Hartford Corridor exceeded 30°C. The 1 h standard was exceeded at the Bethlehem, Charleroi and Seven Corners sites. The forecasting technique correctly identified the Charleroi and Seven Corners sites to reach or exceed the standard, but incorrectly identified the York site to exceed the standard. The predicted maximum concentration at York was only 4 $\mu g\ m^{-3}$ above the standard, however.

The next day, maximum surface temperatures decreased ~10°C throughout much of the Northeast as the center of a high pressure system moved across Maine from Canada. Meanwhile, the cold front in the region became stationary over Virginia. The forecasting technique identified the Charleroi site to exceed the standard, but no site actually did. The 1 h average ozone concentration at Charleroi peaked at 206 $\mu g\ m^3$, however, and was the highest concentration in the Northeast.

5. Summary and conclusions

Without a doubt, there is a meteorological dependency on the ozone concentrations at any site. An attempt was made to define the pertinent meteorological variables and to develop a forecasting technique to predict the daily maximum 1 h average ozone concentration at 27 monitoring sites in the northeastern United States. As a result, multiple lin-

ear regression equations relating the maximum 1 h average ozone concentrations to prognostic and climatological meteorological variables and air quality data were developed for each of the 27 sites. Air quality and meteorological data corresponding to 244 randomly selected days in June, July, August and September of 1975, 1976 and 1977 were used as the developmental or dependent data set. Of the dozens of meteorological variables considered in the regression analysis, most were prognostic variables produced by the National Weather Service Limited-Area Fine Mesh Model. The variables statistically selected most often in the regression equations were the next day's maximum surface temperature, the "quadrant of backward trajectory approach", the sea-level pressure at 0000 GMT the next day, and the boundary-layer u wind component at 0600 GMT the next day.

Air quality and meteorological data corresponding to the remaining 122 days of these months were used as the testing or independent data set to calculate statistical parameters to assess the performance of the regression equations as a forecasting technique. The results of the statistical assessment indicated that the forecasts from the regression equations were mediocre, from the standpoint that the average linear correlation coefficient between the observed and the predicted maximum 1 h average ozone concentration was only 0.64, and the average rms error averaged 62 $\mu g\ m^{-3}$. The performance of the persistence forecasts was worse: The linear correlation coefficient was 0.47 while the rms error was 71 $\mu g\ m^{-3}$.

To provide another means of assessing the performance of the regression forecasts, a frequency plot of normalized residuals was constructed and a χ^2 test was conducted using three-by-three contingency tables. The frequency plot of normalized residuals indicated that: 1) there was a slightly greater chance of underpredicting the daily maximum 1 h average ozone concentrations; 2) nearly half of the predictions were within 20% of the observations; and 3) 77% of the predictions were within 40% of the observations. The χ^2 test indicated that 1) at each monitoring site, the regression equations performed significantly better than chance and 2) at all but four of the monitoring sites, the regression forecasts were better than the persistence forecasts.

Finally, a series of maps was presented comparing areas of the northeastern United States where the daily maximum 1 h average ozone concentrations exceeding the standard were observed and predicted by the regression equations during an ozone episode of June 1975. Based on this series and others not presented in this paper, the observed and predicted areas agreed well preceeding, during and succeeding ozone episodes. This suggested that the regression equations accounted for those meteorological scenarios responsible for the episodic occurrences.

The results of this study indicate that prognostic and climatological meteorological variables alone accounted for much of the day-to-day and site-to-site variations of the daily maximum 1 h average ozone concentrations. There were factors that, from a physical perspective, influence the ozone concentrations, but were not considered in the regression analysis (e.g., precursor emissions, ozone depletion, vertical mixing). The omission of these factors contributed to the errors in the predictions. The fact that the prognostic meteorological variables in the regression equations did not always verify, also contributed to the errors. Although the predicted maximum 1 h average ozone concentrations often differed by more than 20% of the observation, the results indicated that the regression equations performed well in predicting the trends and patterns of the maximum 1 h average ozone concentrations.

REFERENCES

Bennett, A. F., and L. M. Leslie, 1979: Statistical correction of dynamic prognoses in the Australian region. *Mon. Wea. Rev.*, **107,** 1254–1262.

Dixon, W. J., 1977: *BMDP Biomedical Computer Programs: P. Series.* University of California Press, 880 pp.

Glahn, H. R., and R. A. Allen, 1966: A note concerning the "inflation" of regression forecasts. *J. Appl. Meteor.*, **5,** 124–126.

——, and D. A. Lowry, 1972: The use of Model Output Statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11,** 1203–1211.

Grant, A. M., 1956: The application of correlation and regression to forecasting. Meteor. Study 7, Bureau of Meteor., Melbourne, Australia, 21 pp.

Guldberg, P. H., R. D. Siegel, R. B. D'Agostino and G. L. Gipson, 1977: Air quality impact of the energy shortage. *J. Appl. Meteor.*, **16,** 3–10.

Harnack, R. P., 1979: A further assessment of winter temperature predictions using objective methods. *Mon. Wea. Rev.*, **107,** 250–267.

Heffter, J. L., and A. D. Taylor, 1975: Trajectory model, Part I. A regional-continental scale transport, diffusion, and deposition model. NOAA Tech. Memo. ERL ARL-50, 28 pp.

Karl, T. R., 1979: Potential application of Model Output Statistics (MOS) to forecasts of surface ozone concentrations. *J. Appl. Meteor.*, **18,** 254–265.

Klein, W. H., 1963: Specification of precipitation from the 700 mb circulation. *Mon. Wea. Rev.*, **91,** 527–536.

——, B. M. Lewis and C. W. Crockett, 1962: Objective forecasts of daily and mean surface temperature. *Mon. Wea. Rev.*, **90,** 11–17.

Meisner, B. N., 1979: Ridge regression—time extrapolation applied to Hawaiian rainfall normals. *J. Appl. Meteor.*, **18,** 904–912.

Paegle, J. N., 1973: Prediction of precipitation probabilities based on 500 mb flow types. *J. Appl. Meteor.*, **13,** 213–220.

Panofsky, H. A., and G. W. Brier, 1965: *Some Applications of Statistics to Meteorology.* Pennsylvania State University Press, 224 pp.

Russo, J. A., Jr., I. Enger and E. L. Sorenson, 1964: A statistical approach to the short-period prediction of surface winds. *J. Appl. Meteor.*, **3,** 126–131.

Wolff, G. T., and P. J. Lioy, 1978: An empirical model for forecasting maximum daily ozone levels in the northeastern U.S. *J. Air Pollut. Control Assoc.*, **28,** 1034–1038.