

Ozone Forecasting in Philadelphia – 2006
Final Report

Prepared for the
Delaware Valley Regional Planning Commission

William F. Ryan
Amy K. Huff
Department of Meteorology
The Pennsylvania State University
University Park, PA 16802

October 31, 2006

Executive Summary

1. Daily ozone (O₃) forecasts for the Philadelphia non-attainment area, including southern New Jersey, northern Delaware and extreme northeastern Maryland, were provided for the period May 2, 2006 to September 7, 2005. Two Code Red and 17 Code Orange cases occurred during 2006. This is about normal for the 2003-2006 period but much less than the earlier 1994-2002 period.

2. The summer of 2006 was very warm and wet overall with portions of the forecast area observing temperatures in the highest 95th percentile of the past 30 years (1971-2000). The precipitation data is somewhat misleading as the excess for the summer months (4.65") is due entirely to a 4.7" surplus during a rainy month of June. The months of July and August were warmer than normal with average precipitation.

3. The correct AQI forecast color code was issued in 70% of all cases. Of the missed code forecasts, the vast majority (73%) were errors between Code Green and Code Yellow. This performance was less than the 2003-2005 average correct code percentage of 76%.

4. The forecasts were accurate for the high O₃ cases (Code Orange or Code Red). Code Orange or Code Red was *observed* on 19 cases with 13 of these cases (68%) carrying forecasts of Code Orange or higher. As has been the case in previous forecast seasons, Code Orange or higher *forecasts* (Ozone Action Days) were also accurate. In 2006, 16 Ozone Action Days were forecast with 13 (81%) observing Code Orange or higher.

5. The median absolute error (MAE) of the numeric O₃ forecasts was 5.0 parts per billion by volume (ppbv) or within 6.3% of observed O₃. This was an improvement over the average MAE for the previous three seasons of 6.6 ppbv. The forecasts showed a large increase in skill over standard benchmark measures – a 44% skill improvement over persistence (the previous day's observations used as the current day forecast).

6. As part of a cooperative agreement, the National Weather Service and the U.S. EPA have developed a coupled chemistry-weather model to predict O₃. For these cases, the NOAA model outperformed the best regression-based forecast guidance for all standard error measures. In the observed Code Orange cases, the NOAA model predicted Code Orange in 14 of 17 cases at the cost of only three false alarms during the season.

7. Recent O₃ observations continue to suggest that regional NO_x controls ("NO_x SIP Rule") are working to reduce peak O₃ in Philadelphia. For the period 2003-2006, mean peak O₃ concentrations and the frequency and magnitude of extreme O₃ cases have been reduced over the previous 9 year period.

Overview: Weather and Ozone 2006

Daily ozone forecasts for the Philadelphia non-attainment area, including southern New Jersey, northern Delaware and extreme northeastern Maryland, were provided for the period May 2, 2006 to September 7, 2006. A time series of daily peak O₃ and forecast O₃ for this period ([Figure 1](#)) shows the usual large oscillations in daily O₃ along with intermittent multi-day periods of persistent high and low concentrations. These multi-day events are driven by synoptic scale (2-5 day) weather patterns. Multi-day high O₃ episodes occurred throughout the summer but were concentrated during the period from late June to early August – this is the typical time for high O₃ to occur in Philadelphia. Two Code Red and 17 Code Orange cases occurred during 2006 ([Figure 2](#)).

The summer of 2006 was very warm and wet overall. Differences from normal in precipitation and temperature measured at Philadelphia International Airport (PHL) are shown in [Figure 3](#). Warmer than normal weather was observed across the region during this period ([Figure 4](#)) with portions of the forecast area observing temperatures in the highest 95th percentile of the past 30 years (1971-2000). The overall statistics are a bit misleading with respect to precipitation. The excess for the summer months (4.65”) is due entirely to a 4.7” surplus during a rainy month of June. The months of July and August were warmer than normal with average precipitation. The regional precipitation maps show this effect with observed precipitation amounts that ranked in the wettest 95th percentile for the June-August period ([Figure 5](#)). However, if only the months of July and August are ranked, then the region observed near normal or slightly dry conditions ([Figure 6](#)).

It is well known that O₃ is enhanced during hot weather ([Figure 7](#)). This relationship is a function of the temperature dependence of O₃ precursor emissions as well as the likelihood of increased UV radiation – more efficient photochemistry – on hot days. Hot ($\geq 90^\circ$ F) days were quite frequent in 2006 (27 days) compared to a longer term average (1997-2005) of 24.8 days. Overall, for the period 1997-2006, 75% of all days with maximum temperature $\geq 90^\circ$ F also reach the Code Orange (≥ 85 ppbv for an 8-hour average) threshold although a much lower percentage (26%) reach Code Red.

Observed O₃ and Forecast Performance

Observed O₃

Daily mean peak O₃ for the summer of 2006 was 63.8 ppv. This is about average for the most recent period (2003-2005) but is below the longer term average (1994-2002) of 69.9 ppbv ([Figure 8](#)). There were 19 days above the Code Orange threshold. Again, this is consistent with recent (2003-2005) average of 17.5 days and well below the longer term (1994-2002) average of 36 days. There were 6 multi-day episodes of Code Orange O₃ concentrations: May 29-June 1, June 17-19, July 2-3, July 11-12, July 17-19 and July

31-August 3. Analyses of these multi-day episodes will be contained in a web site to be prepared later.

Color Code Forecasts

During the 2006 summer O₃ season forecasts were issued daily for the period from early May to early September. The forecasts were issued as a color code to the public with numerical (ppbv) forecasts prepared internally as a tool for quantitative determination of forecast skill. The correct color code was issued in 70% of all cases ([Table 1](#)). Of the missed code forecasts, the vast majority (73%) were errors between Code Green and Code Yellow. This performance was less than the 2003-2005 correct code percentages of 76%.

For the high O₃ cases (Code Orange or Code Red), the forecasts were accurate. Code Orange or Code Red was *observed* on 19 cases with 13 of these cases (68%) carrying forecasts of Code Orange or higher. As has been the case in previous forecast seasons, Code Orange or higher *forecasts* (Ozone Action Days) were accurate. In 2006, 16 Ozone Action Days were forecast with 13 (81%) observing Code Orange or higher.

ppbv Forecasts

For the 2006 season, the forecasts had a high bias (+2.6 ppbv). The median absolute error (MAE) for the consensus (public) forecasts was 5.0 ppbv. This was an improvement over the previous three-season average of 6.6 ppbv. Forecasts prior to 2003 were issued for peak 1-hour O₃ and so are not directly comparable. The root mean square (rms) error, a measure of forecast consistency, was 11.3 ppbv which is higher than the 2003-2005 average of 10.0 ppbv. The forecasts showed a large increase in skill over standard benchmark measures – a 44% skill improvement over persistence (the previous day's observations used as the current day forecast).

Statistical Model Performance.

The forecasts issued to the public are based on statistical forecast models that have been in use for several years. The statistical models are strongly influenced by forecasted temperature and cannot completely account for other meteorological effects (e.g., thunderstorms) that greatly affect O₃ concentrations.

Currently two statistical models are in use in Philadelphia (R0302 and LN2001). Both models use multiple linear regression methods with the key predictors typically including temperature, relative humidity, wind at the surface and aloft, along with previous day peak O₃ and climatology. The correct color code was forecast by the “best” statistical guidance in 68% of all cases. This is lower than the public forecast skill of

70%. The key shortcomings of the statistical model are over-predictions at the good to moderate threshold – moderate forecast while good conditions observed ([Table 2](#)).

Quantitative statistics for the consensus (public) forecast and statistical guidance are given in [Table 3](#). The overall performance for the R0302 model guidance was very close to the public (consensus) forecasts, while the LN2001 model was less skillful. The mean error and rms error was higher for the statistical guidance suggesting inconsistent skill by these methods.

Forecast Performance in High O₃ Cases

The most important measure of forecast skill is performance in the higher end of the O₃ distribution. These cases, involving observations and forecasts of O₃ in the Code Orange (O₃ Action Day) range, have the greatest implications for pollution control strategies and public health.

In [Table 4](#), a summary of forecast performance in *observed* cases of Code Orange or higher O₃ is given. All forecasts, consensus or statistical guidance, showed a low bias in these cases with the highest bias magnitude for the consensus forecast (-4.2 ppbv). For all other quantitative error measures (median absolute error, mean absolute error and rms error), the consensus forecast improved slightly on the statistical model guidance. This has generally been the case over the past forecast seasons. For median absolute error, the expert forecasts improved on the best statistical guidance by 23% with a 14% improvement on the mean and rms error measures.

The complementary measure of forecast skill in high O₃ cases is performance in cases of *forecast* high O₃. These results are shown in [Table 5](#). Only three false alarms (O₃ Action Day issued but not observed) were issued in 2006 and, in all these cases, Code Orange concentrations were observed either on the following day (2 cases) or the preceding day (1 case). In all, 13 of 16 (81%) of the forecasts of O₃ Action Days verified with observed Code Orange O₃. Of the two statistical models in use in 2006, only the R0302 model showed comparable skill in the high ozone cases. The LN2001 model tended to over-predict high O₃ with 8 cases of false alarms. We also tested an average of the R0302 and LN2001 models (“blend”) but it did not improve on the stand alone R0302 model. As a result, we will discard the LN2001 model for future forecasting. Our expectation is to develop a new statistical model using other statistical techniques – Classification and Regression Tree (CART) or neural networks - for 2007.

A variety of forecast skill measures for observed and forecast Code Orange cases are given in [Table 6](#). These scores measure the skill of the forecasts at accurately predicting cases above a threshold concentrations. In this case, the threshold is 85 ppbv for an 8-hour maximum. The skill scores are described in more detail in [Appendix A](#). Overall, the consensus forecast tended to be conservative. That is, few false alarms were issued (FAR = 0.19) but at the cost of missing a few Code Orange cases (POD = 0.68). The bias measure is < 1.00 which suggests a tendency to under-predict at this threshold.

Overall skill is best measured by the Conditional Skill Index (CSI) and Heidke Skill Score (HSS) measures. As Code Orange O₃ is a relatively rare event, the most numerous category is the “correct null” forecast, that is, low O₃ forecast and observed. These forecasts are typically easy to make and not of critical interest. The CSI calculation removes the influence of the correct null forecast. For 2006, the consensus forecast skill score of 0.59 (CSI has a range [-1,1]) is good. The HSS score compares the forecasts to a randomly generated set of forecasts that are constrained to have the same number of cases in each color code as observed. Again, on the range of [-1,1], the HSS of 0.70 is a good result.

Numerical Forecast Guidance

As part of a cooperative agreement, the National Weather Service and the U.S. EPA have developed a coupled chemistry-weather model to predict O₃. (see, http://www.nws.noaa.gov/ost/air_quality/index.htm) This model was tested in 2003 and 2004 and became fully operational in 2005. The NOAA/EPA model has shown reasonable skill overall and was able to predict some of the high O₃ cases in 2005. In 2006, we undertook a more searching analysis of numerical model forecast performance in the Philadelphia area.

Using data provided online by NOAA, we archived peak O₃ forecasts for 9 of the 19 O₃ monitor locations in the Philadelphia metropolitan area. These monitors (Bellefonte, New Garden, Ancora, Rider, Bristol, Norristown, Northeast Philadelphia, Colliers Mills and Center City) were chosen because they are either typically among the highest O₃ monitors in the forecast area and also to give good geographic coverage. For the period May 28-August 31, data was collected for 85 of 106 days. This included 17 of the 19 days on which Code Orange or higher O₃ was observed. Peak O₃ for the nine monitor locations was used as the forecast peak O₃ for the metropolitan area. For these cases, the NOAA model outperformed the R0302 regression guidance by all standard error measures. Forecast bias was +3.1 ppbv (+4.9 ppbv for R0302), median absolute error was 6.0 ppbv (7.7 ppbv for R0302) and mean absolute error was 8.5 ppbv (9.1 ppbv for R0302). In the observed Code Orange cases, the NOAA model predicted Code Orange in 14 of 17 cases at the cost of only three false alarms during the season. While the NOAA model under-predicted peak concentrations in the two Code Red cases, it also correctly predicted concentrations during the heat wave episode of July 31-August 3 where the regression models, due to high temperature forecasts, over-predicted.

The results from 2006 suggest that the NOAA forecast model predictions should be more fully integrated into the forecast process. Further research along these lines will be carried out in 2007.

Effect of NO_x emissions reductions

The NO_x SIP call rule, issued by the EPA in 1998 to reduce regional formation and transport of O₃, was phased in over the eastern US during the period 2003-2005. The main effect of the rule was to reduce power plant NO_x emissions ([Figure 8](#)). Regional, that is, non-urban, concentrations of O₃ are limited chemically by NO_x concentrations. As a result, the pollution controls introduced as part of the SIP call rule are expected to reduce regional (background) O₃. As a significant portion of urban-scale O₃ is due to transported air masses on a regional scale, this lower background O₃ should translate into O₃ reductions in urban areas.

As the NO_x controls have only been in effect beginning in 2003, and 2004 was a cool, low O₃ season, no conclusive results regarding the efficacy of NO_x controls can be determined at this time. But, there is mounting anecdotal evidence that NO_x emissions controls are working to reduce local peak O₃.

On the coarsest scale, mean peak O₃ for the summer season has been reduced since 2002 ([Figure 9](#)). Mean peak daily O₃ for the period 1994-2002 was 69.9 ppbv – with similar values for shorter period of 1997-2002. For period 2003-2006, mean daily peak O₃ is 63.1 ppbv or 10% lower. This is a coarse statistical measure and does not account for differences in meteorology although climate statistics ([Figure 3](#)) shows that the 2003-2006 period, while wetter than normal, has also been warmer than normal.

Lower mean peak O₃ concentrations have translated in changing frequencies of AQI color codes. In 2006, Code Green O₃ occurred most frequently with only 15% of all days reaching Code Orange or higher ([Figure 10](#)). For period 2003-2006, similar results are found with 53% Code Green cases, 32% Code Yellow and 15% Code Orange or higher ([Figure 11](#)). The relative frequencies, particularly at the higher end of the O₃ distribution, are markedly different for the 1994-2002 period ([Figure 12](#)). During that period, the Code Orange or higher percentage was 23%. Similar differences in frequency of occurrence are found for extreme high O₃ cases. The average frequency of days ≥ 115 ppbv was 4.8 in the 1994-2002 period and only 0.8 for the 2003-2006 period ([Figure 13](#)).

The results noted above, of course, do not account for year to year differences in temperature and other O₃-conductive weather features. In [Table 7](#), the percentage of warm ($\geq 90^\circ$ F) days that reach the Code Orange or Code Red threshold are given. For days $\geq 90^\circ$ F, Code Orange or higher O₃ occurred at an 83% rate with 35% of the days reaching Code Red. For the more recent period, 2003-2006, only 60% of these hot cases resulted in Code Orange or higher O₃ and only 10% reached Code Red. This suggests that less O₃ is being observed in the most O₃-conductive cases.

Another way of looking at the temperature-O₃ association is given in [Figure 14](#). In that figure, O₃ concentrations of the 75th percentile of observations for selected maximum temperature ranges are given. For the most recent period, O₃ concentrations are lower, on the order of 5-15 ppbv, across the warm temperature spectrum. The effect of the changes in O₃ concentrations in the forecast area on forecast skill is shown in [Table](#)

[8](#) and [Table 9](#). In general, the main difference is an increase in the high bias (over-prediction) of the forecasts.

Conclusions

Daily ozone (O₃) forecasts for the Philadelphia non-attainment area, including southern New Jersey, northern Delaware and extreme northeastern Maryland, were provided for the period May 2, 2006 to September 7, 2005. Two Code Red and 17 Code Orange cases occurred during 2006. This is about normal for the 2003-2006 period but much less than the earlier 1994-2002 period.

The summer of 2006 was very warm and wet overall with portions of the forecast area observing temperatures in the highest 95th percentile of the past 30 years (1971-2000). The season averaged precipitation data is somewhat misleading as the excess for the summer months (4.65”) is due entirely to a 4.7” surplus during a rainy month of June. The months of July and August were warmer than normal with average precipitation.

The correct AQI forecast color code was issued in 70% of all cases. Of the missed code forecasts, the vast majority (73%) were errors between Code Green and Code Yellow. This performance was less than the 2003-2005 average correct code percentage of 76%.

The forecasts were accurate for the high O₃ cases (Code Orange or Code Red). Code Orange or Code Red was *observed* on 19 cases with 13 of these cases (68%) carrying forecasts of Code Orange or higher. As has been the case in previous forecast seasons, Code Orange or higher *forecasts* (Ozone Action Days) were also accurate. In 2006, 16 Ozone Action Days were forecast with 13 (81%) observing Code Orange or higher.

The median absolute error (MAE) of the numeric O₃ forecasts was 5.0 parts per billion by volume (ppbv) or within 6.3% of observed O₃. This was an improvement over the average MAE for the previous three seasons of 6.6 ppbv. The forecasts showed a large increase in skill over standard benchmark measures – a 44% skill improvement over persistence (the previous day’s observations used as the current day forecast).

As part of a cooperative agreement, the National Weather Service and the U.S. EPA have developed a coupled chemistry-weather model to predict O₃. For these cases, the NOAA model outperformed the best regression-based forecast guidance for all standard error measures. In the observed Code Orange cases, the NOAA model predicted Code Orange in 14 of 17 cases at the cost of only three false alarms during the season.

Color Code Forecasts Philadelphia Forecast -2006				
Observed				
Forecast	Green	Yellow	Orange	Red
Green	55	13	0	0
Yellow	14	23	6	0
Orange	0	3	10	1
Red	0	0	1	1

Table 1. Color code forecasts for the Philadelphia forecast area – 2006.

Statistical Forecast Guidance Philadelphia Forecast -2006				
Observed				
Forecast	Green	Yellow	Orange	Red
Green	49	8	0	0
Yellow	30	27	5	0
Orange	0	4	10	1
Red	0	0	1	1

Table 2. As in Table 1 but for the “best” statistical forecast guidance model.

Summary Statistics Forecasts and Statistical Guidance Philadelphia Forecast-2006 (all measures in ppbv)						
	Obs	Forecast	R0302	LN2001	Blend	Persistence
Mean	63.8	66.4	67.7	70.2	69.0	64.1
Bias		+2.6	+3.9	+6.4	+5.2	+0.3
Mean AE		6.9	6.8	7.8	7.3	11.9
Median AE		5.0	5.6	6.4	5.6	10.0
Rms error		11.3	11.6	12.9	12.1	15.2

Table 3. Summary statistics for consensus forecast, two statistical models (R0302 and LN2001), an average of the two models (Blend) and the reference persistence forecast.

Philadelphia Forecast Area -2006 Observed High O₃ Cases						
	Observed	Forecast	R0302	LN2001	Blend	Persistence
Median (ppbv)	90	89	89.1	90.4	89.8	86.0
Mean (ppbv)	94.4	90.2	91.5	93.4	92.4	86.0
Bias (ppbv)		-4.2	-2.9	-1.0	-2.0	-8.4
Median AE		5.0	8.0	6.5	7.1	12.0
Mean AE (ppbv)		8.5	9.9	9.9	9.8	14.5
rms error (ppbv)		11.4	13.4	13.4	13.3	19.5
Orange Fcst/Obs		13/19	14/19	15/19	14/19	11/19
≥82 ppbv Fcst		16	15	16	16	11

Table 4. Summary forecast statistics for consensus and statistical model guidance for cases in which observed peak O₃ in the Philadelphia area reached or exceeded the Code Orange AQI threshold (85 ppbv for an 8-hour average).

Philadelphia Forecast Area -2006 Forecast High O₃ Cases					
	Forecast	R0302	LN2001	Blend	Persistence
Median (ppbv)	90.0	90.4	90.4	91.3	90.0
Mean (ppbv)	94.0	94.9	95.0	95.5	94.4
Bias (ppbv)	+3.3	+3.6	+7.8	+5.7	+14.1
Median AE	5.5	9.0	10.4	10.2	13.0
Mean AE (ppbv)	8.7	8.9	11.3	10.0	16.8
rms error (ppbv)	11.7	11.8	14.1	12.9	20.8
Orange Fcst/Obs	13/16	14/18	15/23	14/19	10/19

Table 5. Summary forecast statistics for consensus and statistical model guidance for cases in which peak O₃ in the Philadelphia area was forecast to reach or exceed the Code Orange AQI threshold (85 ppbv for an 8-hour average).

Skill Scores Philadelphia Forecast Area - 2006					
	Fcst	R0302	LN2001	Persistence	Blend
POD	0.68	0.74	0.79	0.58	0.74
Miss	0.32	0.26	0.21	0.42	0.26
FAR	0.19	0.22	0.35	0.66	0.26
Accuracy	0.93	0.93	0.93	0.87	0.92
CNull	0.95	0.95	0.96	0.93	0.93
Bias	0.84	0.95	1.21	1.00	1.00
CSI	0.59	0.61	0.56	0.41	0.58
TSS	0.66	0.70	0.72	0.50	0.69
HSS	0.70	0.72	0.66	0.50	0.69

Table 6. Skill score measures for the consensus forecast, statistical guidance and reference (persistence) forecasts. All measures are detailed in *Appendix A*

POD = Probability of Detection

Miss = Miss Rate

FAR = False Alarm Rate

CNull = Correct Null Forecast

CSI = Critical Success Index, or Threat Score

TSS = True Skill Statistic, also known as Peirce score, Hanssen-Kuipers discriminant

HSS = Heidke Skill Score

Philadelphia Forecast Area Days $\geq 90^{\circ}\text{F}$ and O_3 Concentration Threshold		
	Number of Cases	Percent Above Threshold
1997-2006		
≥ 85 ppbv	187	75%
≥ 105 ppbv	66	26%
1997-2002		
≥ 85 ppbv	134	83%
≥ 105 ppbv	57	35%
2003-2006		
≥ 85 ppbv	53	60%
≥ 105 ppbv	9	10%

Table 7. Frequency of 8-hour peak O_3 above the Code Orange and Code Red thresholds for high temperature cases.

Summary Forecast Statistics – Consensus Forecasts Philadelphia Forecast Area				
	2006	2005	2004	2003
Bias (ppbv)	+2.6	+2.2	+1.6	+0.4
Mean AE (ppbv)	6.9	8.8	7.0	7.3
Median AE (ppbv)	5.0	7.0	6.0	7.0
rms error (ppbv)	11.3	12.1	8.8	9.0
Days \geq 90 F (JJA)	23	7	27	25
Days \geq 85 ppbv	20	10	20	19

Table 8. Consensus forecast results for the 2003-2006 seasons.

Summary Forecast Statistics – Statistical Guidance Philadelphia Forecast Area				
	2006	2005	2004	2003
Bias (ppbv)	+3.9	+3.9	+1.7	+1.3
Mean AE (ppbv)	6.8	10.6	7.6	7.6
Median AE (ppbv)	5.0	7.0	6.0	6.4
rms error (ppbv)	11.6	13.0	10.1	9.6
Days \geq 90 F (JJA)	23	7	27	25
Days \geq 85 ppbv	20	10	20	19

Table 9. As in Table 8 but for regression forecast guidance.

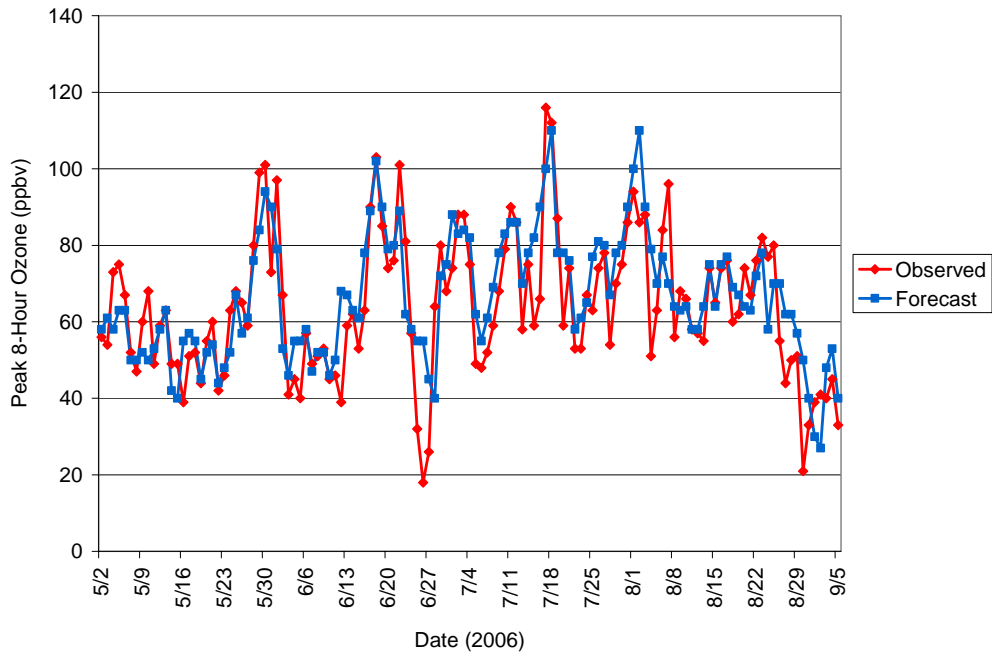


Figure 1. Time series of daily peak 8-hour O₃ in the Philadelphia metropolitan area and forecast peak O₃.

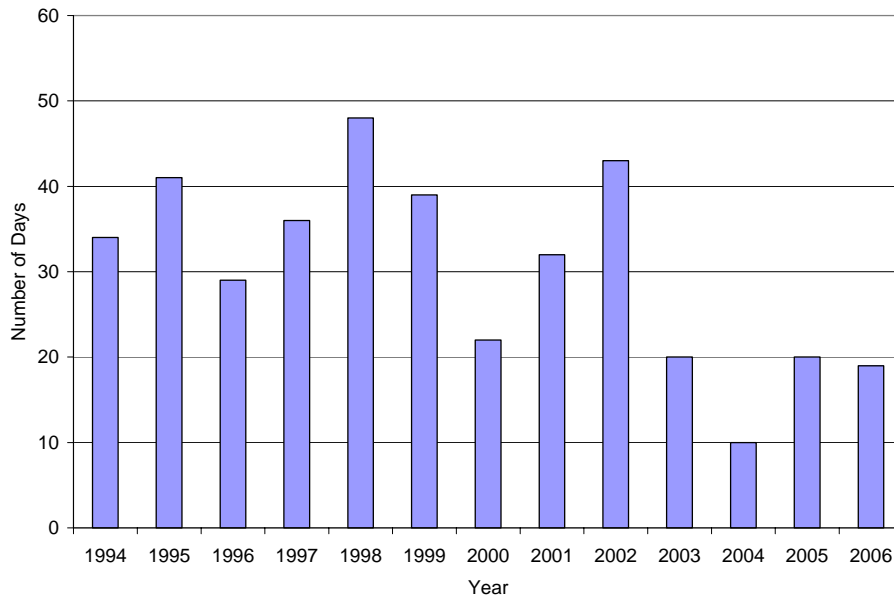


Figure 2. Number of days in Philadelphia during the 2006 summer forecast season exceeding the Code Orange threshold (85 ppbv, 8-hour average). For the period 1994-2002, the average number of days above this threshold is 36, for the more recent period, 2003-2006, the average is 17.2.

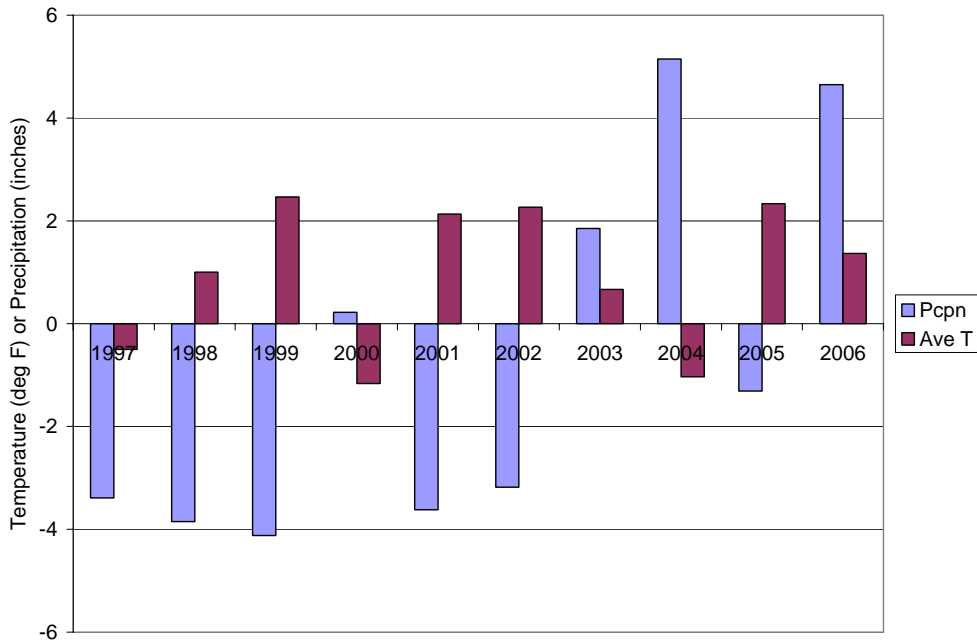


Figure 3. Difference from normal in precipitation for the June-August period for Philadelphia International Airport (PHL) (blue column) and difference from normal in mean temperature (red column).

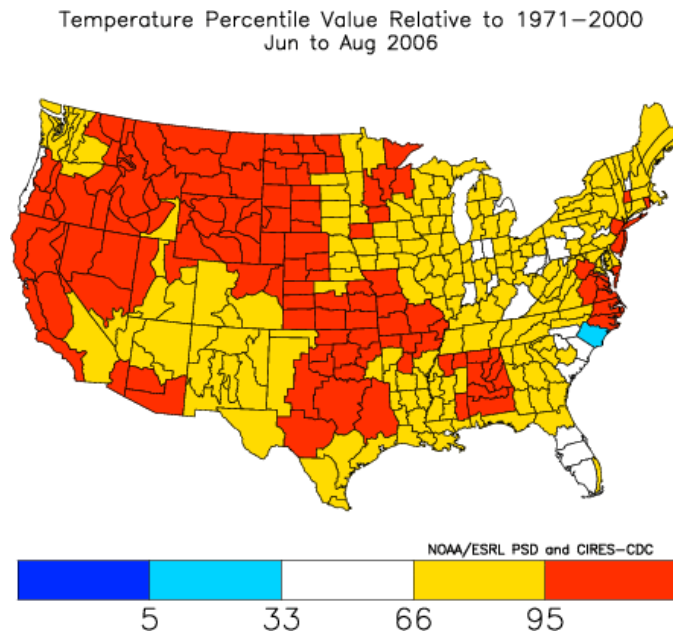


Figure 4. Mean temperature for June-August, 2006 expressed as a percentile compared to the 1971-2000 period. Figure courtesy of NOAA Climate Diagnostics Center.

Precipitation Percentile Value Relative to 1971–2000
Jun to Aug 2006

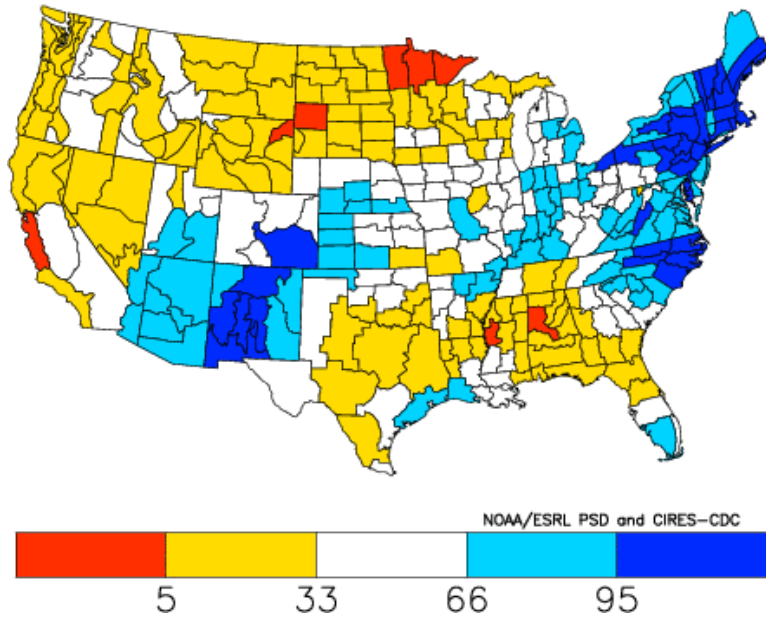


Figure 5. As in Figure 4 but for precipitation.

Precipitation Percentile Value Relative to 1971–2000
Jul to Aug 2006

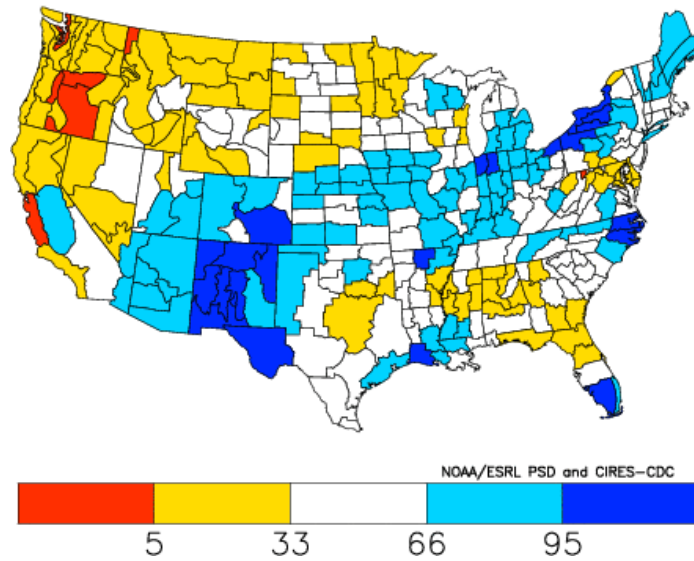


Figure 6. As in Figure 5 but for only the months of July and August, 2006.

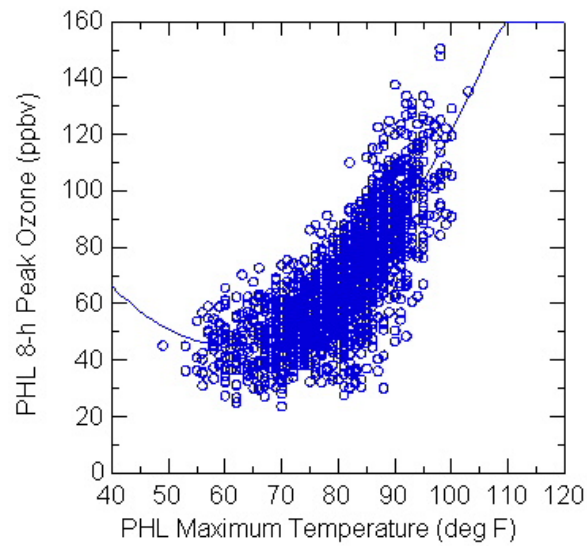


Figure 7. Maximum surface temperature measured at PHL and Philadelphia non-attainment area peak 8-hour O₃ concentrations for the period May-September, 1994-2004. Best quadratic fit line is included.

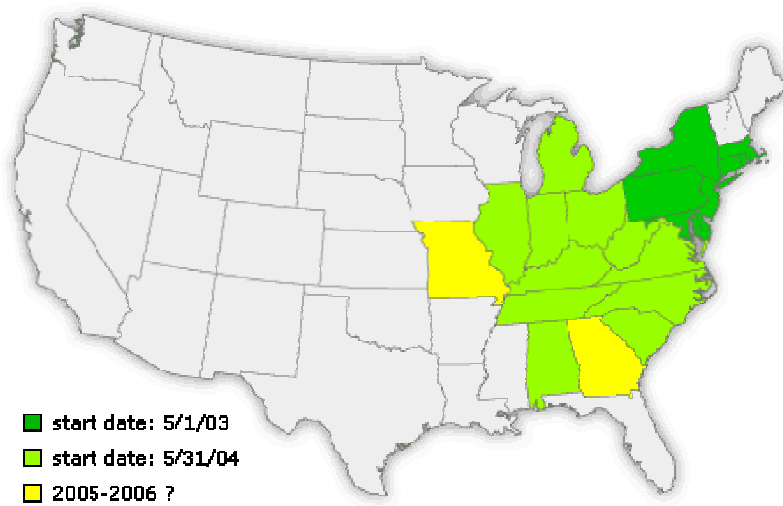


Figure 8. Regions affected by the NO_x SIP call rule.

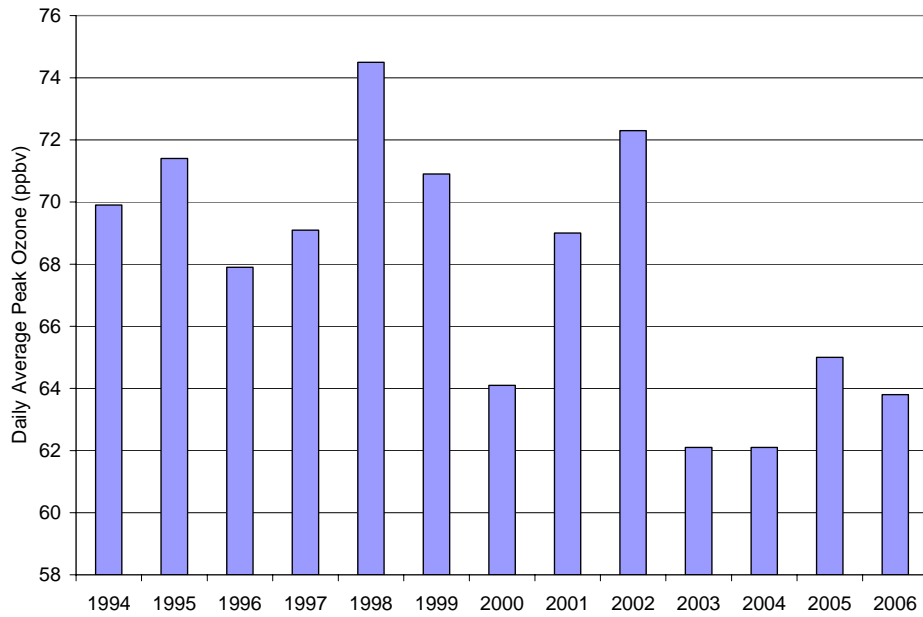


Figure 9. Mean peak daily O₃ for the Philadelphia forecast area.

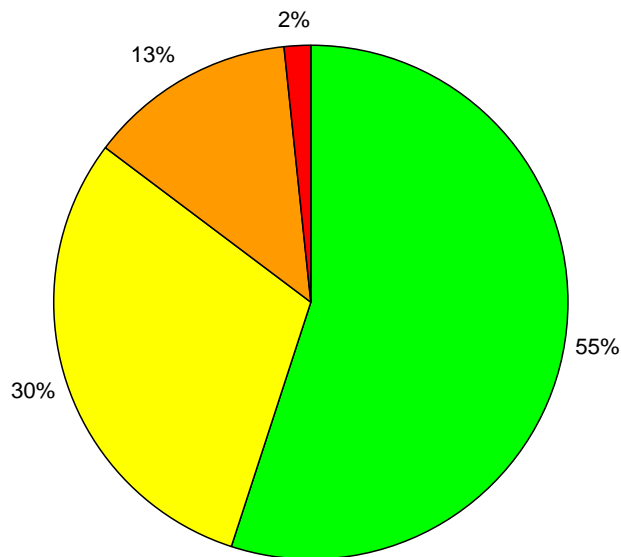


Figure 10. Frequency of occurrence of ozone concentrations in the Philadelphia forecast area by color code for 2006.

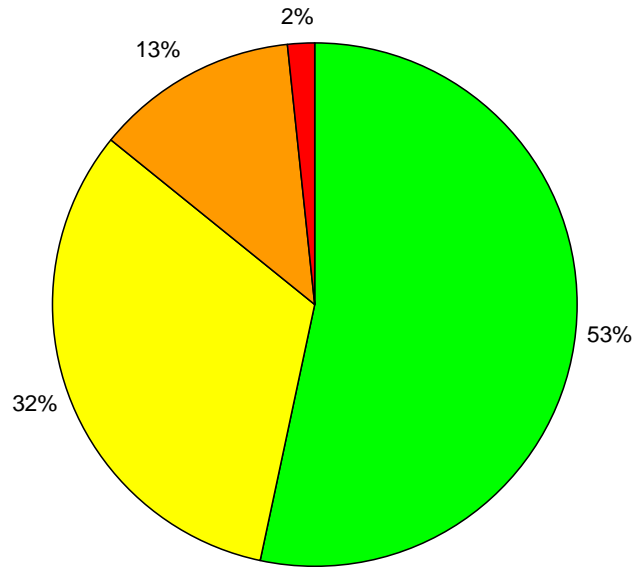


Figure 11. As in Figure 10, but for the period 2003-2006.

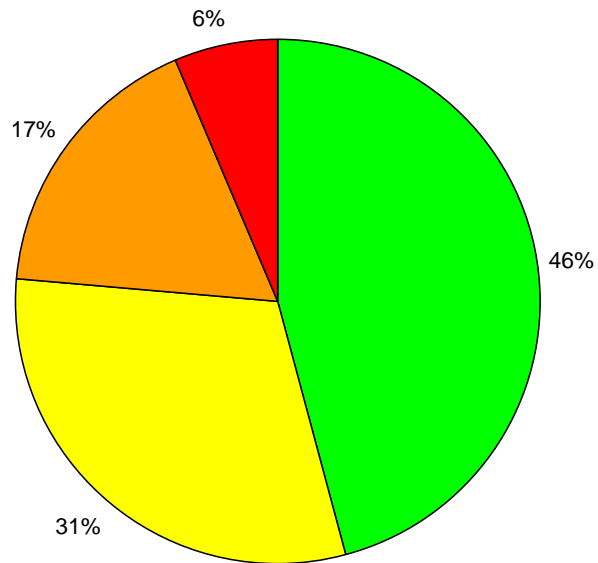


Figure 12. As in Figure 10, but for the period 1994-2002.

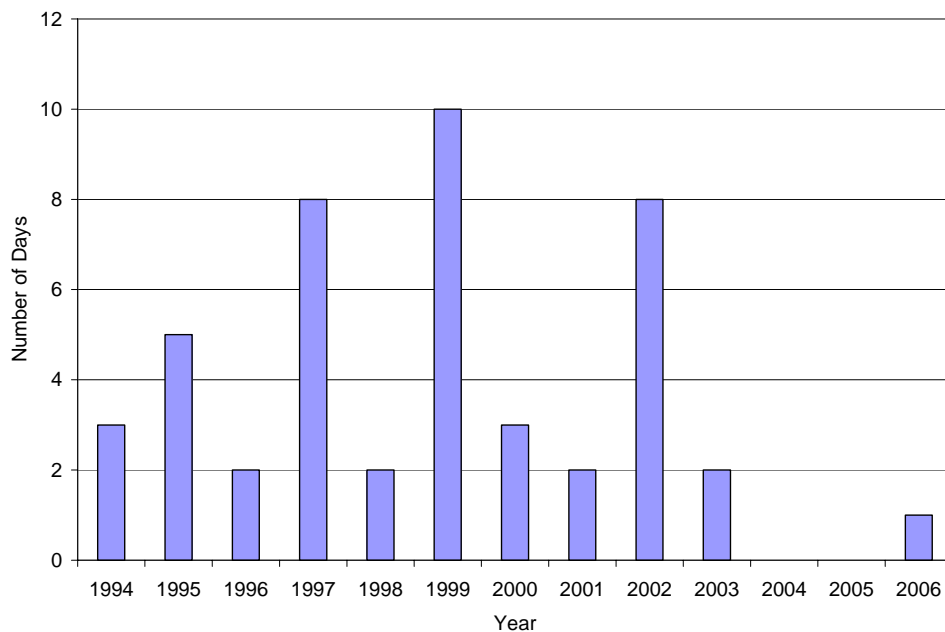


Figure 13. Number of days with peak 8-hour O₃ concentrations ≥ 115 ppbv in the Philadelphia forecast area. The average for the period 1994-12002 is 4.8 days, for 2003-2006, 0.75 days.

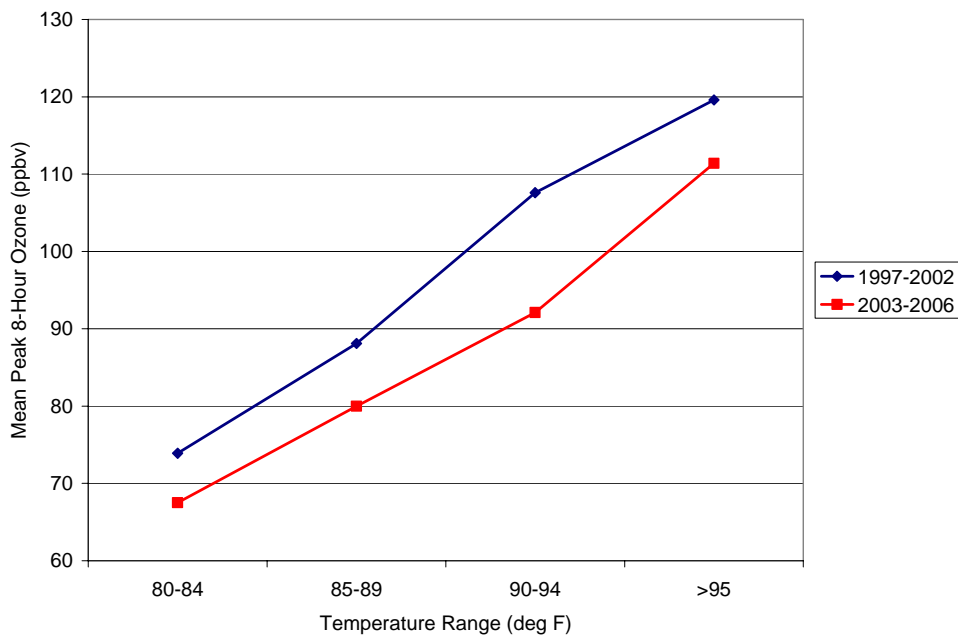


Figure 14. The 75th percentile of observed O₃ concentrations for selected maximum temperature bins in the Philadelphia forecast area.

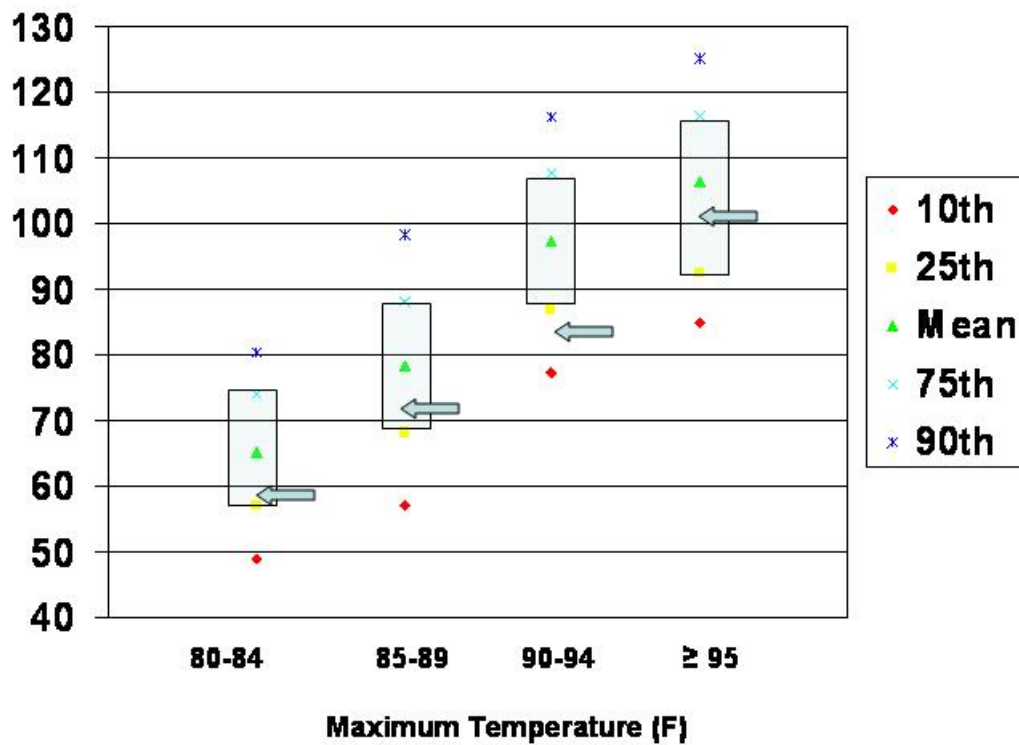


Figure 15. Distribution of O₃ concentrations for selected maximum temperature bins. For the period 1994-2002, the 10th to 90th percentile O₃ concentrations are given. The light blue arrows show the mean O₃ concentrations for the same bins for the period 2003-2006.

Appendix A. Skill of Threshold Forecasts

The determination of the skill of a threshold forecast begins with the creation of a contingency table of the form:

Contingency Table for Threshold Forecasts			
		Forecast	
		Yes	No
Observed	Yes	a	b
	No	c	d

For example, if Code Red O₃ concentrations are both observed and forecast (“hit”), then one unit is added to “a”. A “false alarm”, Code Red forecast but not observed, is added to “c”.

Basic Skill Measures:

A basic set of skill measures are determined and then used as the basis for further analysis.

$$\text{Bias (B)} = \frac{a+b}{a+c} = 0.69$$

Bias determines whether the same *fraction* of events are both forecast and observed. If $B = 1$, then the forecast is unbiased. If $B < 1$ there is a tendency to under-predict and if $B > 1$ there is a tendency to over-predict. In this case $B = 0.69$ which points to a tendency to under-predict Code Red cases.

$$\text{False Alarm Rate (F)} = \frac{b}{b+d} = 0.02$$

This is a measure of the rate at which false alarms (high O₃ forecast but not observed) occur. For OAD forecasts, this rate is extremely low.

$$\text{Hit Rate (H)} = \frac{a}{a+c} = 0.78$$

The hit rate is often called the “probability of detection”

$$\text{Miss Rate} = 1 - H = 0.22$$

Correct null forecasts:

$$\text{Correct Null (CNull)} = \frac{d}{c+d} = 0.95$$

Accuracy:

$$\text{Accuracy (A)} = \frac{a+d}{a+b+c+d} = 0.78$$

Other Measures:

Generalized skill scores (SS_{ref}) measure the improvement of forecasts over some given reference measure. Typically the reference is persistence (current conditions used as forecast for tomorrow) or climatology (historical average conditions). Because O_3 concentrations have a strong seasonal cycle and concentrations have decreased significantly from the late 1980's, climatology in this case is based on a ten-day running average, centered on the day of interest, from the period 1994-2004.

$$\text{Skill Score (SS}_{\text{ref}}) = \left(1 - \frac{MSE}{MSE_{\text{ref}}}\right) * 100\% = 53\%$$

The skill score is typically reported as a percent improvement over the reference forecast.

Additional measures of skill can be determined. The Heidke skill score (HSS) compares the proportion of correct forecasts to a no skill random forecast. That is, each event is forecast randomly but is constrained in that the marginal totals ($a + c$) and ($a + b$) are equal to those in the original verification table.

$$\text{HSS} = \frac{2(ad - bc)}{(a+c)(c+d) + (a+b)(b+d)} = 0.61$$

For this measure, the range is $[-1, 1]$ with a random forecast equal to zero. The Code Red forecast show skill by this measure.

Another alternative is the **critical success index (CSI) or the Gilbert Skill Score (GSS)** also called the **“threat” score**.

$$\text{CSI} = \frac{a}{a+b+c} = \frac{H}{1+B-H} = 0.47$$

For this measure, the range is [0,1]. Since the correct null forecast is excluded, this type of measure is effective for situations like tornado forecasting where the occurrence is difficult to determine due to observing bias, i.e., tornados may occur but not be observed. This can also be the case for air quality forecasting when the monitor network is less dense. Note, however, that the random forecast will have a non-zero skill.

The **Peirce skill score (PSS)**, also known as the “**true skill statistic**” is a measure of skill obtained by the difference between the hit rate and the false alarm rate:

$$\text{PSS} = \frac{ad - bc}{(a + c)(b + d)} = H - F = 0.74$$

The range of this measure is [-1,1]. If the PSS is greater than zero, then the number of hits exceeds the false alarms and the forecast has some skill. Note, however, that if d is large, as it is in this case, the false alarm value (b) is relatively overwhelmed. The advantage of the PSS is that determining the standard error is relatively easy.

References

Stephenson, D. B., Use of the “odds ratio” for diagnosing forecast skill, *Wea. Forecasting*, **15**, 221-232, 2000.

Wilks, D. S., *Statistical Methods in the Atmospheric Sciences*, Academic Press, 467pp., 1995.