State of Delaware
Final Report: Ozone Observations and Forecasts in 2013

A Report Prepared for the Delaware Department of Natural Resources and Environment

William F. Ryan
Dr. Amy K. Huff
Department of Meteorology
The Pennsylvania State University
REVISED June 2014

Table of Contents

Executive Summary 4
Ozone Observations in 2013 4
Meteorology and Code Orange O_3 Concentrations in 2013 5
Skill of Ozone Forecasts in 2013 6
Performance of Ozone Numerical Forecast Models in 2013 7
Outlook for 2014 8
Appendix A. Skill Measures for Threshold Forecasts 19

Table of Figures

Figure 1. Seasonal (May-September) mean and median peak 8-hour observed O_3 in Delaware for 1990-2013. 9
Figure 2. Daily time series of peak 8-hour observed O_3 in Delaware for 2013 compared to the 2003-2012 average. 9
Figure 3. Daily time series of the difference between peak 8-hour observed O_3 for 2013 compared to the 2003-2012 average. 10
Figure 4. Frequency of Code Purple, Code Red, and Code Orange observed O_3 days in Delaware for 1990-2013. 10
Figure 5. Percent reduction in the occurrence of Code Orange, Red, and Purple observed O_3 days in Delaware since 2002 11
Figure 6. Frequency of AQI color codes for observed O_3 in Delaware for 1990-2002. 11
Figure 7. Frequency of AQI color codes for observed O_3 in Delaware for 2003-2012. 12
Figure 8. Frequency of AQI color codes for observed O_3 in Delaware for 2013. 12
Figure 9. Percent of days exceeding the observed O_3 Code Orange threshold in Delaware for bins of maximum surface temperature 13
Figure 10. Temperature anomalies (in °F) in the U.S. for June-August 2013 compared to the 1981-2010 average 13
Figure 11. Precipitation anomalies (in inches) in the U.S. for June-August 2013 compared to the 1981-2010 average 14
Figure 12. Vector wind anomalies at 850 mb (~1500 m AGL) in the U.S. for June-August 2013. 14
Figure 13. Number of days with peak 8-hour observed O_3 above the 60 ppbv threshold at Shenandoah National Park, VA for June-August of 2004-2013. 15
Figure 14. Peak 8-hour O_3 forecasts and observations for Delaware during May-September, 2013. 15
Figure 15. Error statistics for all peak 8-hour O_3 forecasts in Delaware for 2011-2013. 16
Figure 16. False alarm rate and hit rate for Code Orange O_3 forecasts in Delaware for 2011-2013. 16
Figure 17. Heidke skill scores and CSI (threat scores) for Code Orange O_3 forecasts in Delaware for 2011-2013. 17
Figure 18. Error statistics for air quality numerical forecast model guidance for all peak 8-hour O_3 predictions in Delaware in 2013. 17

Figure 19. False alarm rate and hit rate for Code Orange O_3 predictions by air quality numerical forecast models in Delaware in 2013. 18
Figure 20. Heidke skill scores and CSI (threat scores) for Code Orange O_3 predictions by air quality numerical forecast models in Delaware in 2013. 18

Executive Summary

- The O_3 season of 2013 was the cleanest on record for the State of Delaware. Only 2 days reached the Code Orange threshold and there were no Code Red or Code Purple days.
- The clean conditions during the summer of 2013 are attributed to continued reductions in O_3 precursor emissions, above average rainfall, and an atypical weather pattern that limited the frequency of westerly transport, which is historically associated with poor air quality days in the Mid-Atlantic region.
- Overall forecast skill (all days) was similar to recent years, although absolute errors were slightly higher due to the challenge of forecasting during the unusually clean O_3 season, with a median absolute error of 6.5 ppbv.
- As a result of the unusually clean O_3 season, forecast skill for the Code Orange O_3 cases was poor. The rate of false alarms was much higher than the recent average, although the total number of false alarms in 2013 (5) was less than in 2011 (10) and 2012 (6). In addition, the forecast hit rate – Code Orange forecast and observed – was considerably below the recent average.
- The numerical air quality forecast models that provide forecast guidance showed a propensity to over-predict peak O_3 in all cases, with a high false alarm rate for Code Orange O_3 cases. The exception was the NOAA-EPA model, which had a lower false alarm rate relative to the other available operational air quality models for Delaware in 2013.
- In the absence of forecast guidance from the SUNY numerical forecast model in 2013, the NCDENR model was added to the "ensemble" of forecast models. The NCDENR model proved to be useful and will be included in the 2014 ensemble.
- Re-introducing purely statistical forecast guidance will be investigated in 2014. Due to significant reductions in O_3 precursor emission beginning in 2002, older statistical guidance performed poorly, and the statistical model based on historic O_3 observations was discontinued in 2012. Because emissions changes have been less abrupt since 2008, a statistical model using 2008-2012 observations will be developed and tested in 2014.

Ozone Observations in 2013

The ozone (O_3) season of 2013 was the cleanest on record for the State of Delaware. As shown in Figure 1, the 2013 seasonal (May 1-September 30) mean and median peak 8-hour observed O_3 concentrations were the lowest since 1990. This can at least partly be attributed to continued regional reductions in O_3 precursor emissions. O_3 concentrations in Delaware have decreased significantly since the regional NO_x emissions reductions regulations, the "NO_x SIP Call," went into effect circa 2002. This decrease is illustrated by a downward trend in the statewide mean and median peak 8-hour O_3 concentrations since 2003 (Figure 1). The daily time series of peak 8-hour observed O_3 in Delaware during the summer of 2013 (May-September) compared to the 2003-2012 average (Figure 2) further illustrates that 2013 was historically clean. The difference between daily peak 8-hour observed O_3 concentrations in 2013 and the 2003-2012 average (Figure 3) shows that daily peak 8-hour observed O_3 concentrations in 2013 fell below the 2003-2012 average on most days.

The 2013 O_3 season was also remarkable for the low occurrence of high O_3 days. During the 2013 O_3 season, there were only 2 days with peak 8-hour observed O_3 in excess of the Code Orange threshold of 76 ppbv, no days in exceedance of the Code Red threshold of 96 ppbv, and no days in exceedence of the Code Purple threshold of 116 ppbv (Figure 4). This marks 2013 as the year with the lowest occurrence of high O_3 days since 1990. Compared to the post-NO_x SIP Call period (2003-2013), Code Orange O_3 days in 2013 were well below the mean frequency of 17 days, and Code Red O_3 days were below the average of 1.2 days per year. Continuing the recent trend, there were no truly extreme O_3 days (8-hour $O_3 \geq 105$ ppbv) in 2013. The last year that 8-hour $O_3 \geq 105$ ppbv was observed was 2012, and the last year that 8-hour $O_3 \geq 115$ ppbv was observed was 2007. This trend is illustrated by Figure 5, which shows that since 2002, observed Code Orange O_3 days have decreased by 54%, observed Code Red O_3 days have decreased by 86%, and observed Code Purple O_3 days have decreased by 88% in Delaware. Pie charts showing changes in the frequency of all observed AQI color codes since 1990 are given in Figures 6-8. These plots further demonstrate the trend toward cleaner O_3 air quality in Delaware since the NO_x SIP Call went into effect. Of particular note is the relatively high proportion of Code Green days (78%) observed in 2013 (Figure 8), which serves to underscore the very clean O_3 conditions that occurred.

Meteorology and Code Orange O_3 Concentrations in 2013

The 2 observed Code Orange O_3 days in 2013 bookended a single multi-day event which occurred on July 16-19. This period coincided with a heat wave, with observed maximum air temperatures ranging from 93-97 °F and no measured precipitation. These conditions are historically conducive for high O_3 concentrations. Temperature and O_3 are strongly correlated overall, but the association of hot weather with Code Orange or higher O_3 has been steadily weakening since 2002, making Code Orange days more difficult to forecast. The percent of days exceeding the Code Orange threshold for a given range of maximum surface temperature is presented in Figure 9. While days with maximum surface temperature ≥ 95° F are still very likely to be poor air quality days, the frequency of Code Orange cases on slightly less hot days (90-94 °F) has dropped from 70% prior to 2003 to less than 10% in 2013. In 2013, which was an usually clean year for O_3, only one-ninth of the days with maximum temperature ≥ 90° F reached the Code Orange threshold.

Despite the weakening effect of hot weather and high O_3, the cause of the historically low O_3 concentrations observed in 2013 does not appear to be related to the occurrence of hot days. Figure 10 indicates that temperatures in Delaware during June-August 2013 were about 1 °F higher than the long-term average. Thus, based on observed air temperatures alone, 2013 should have been an average or slightly above average O_3 season. Figure 11 shows that approximately 6-10 inches more precipitation than the long-term average fell in Delaware during June-August 2013, however. Cloud cover associated with precipitation blocks sunlight and inhibits O_3 formation. Therefore, while summer 2013 was slightly hotter than normal, it was also wetter than normal, which was likely one contributing factor to the historically low observed O_3 concentrations in 2013.

Another key factor in limiting peak O_3 concentrations during summer 2013 appeared to be the predominant transport pattern aloft. During summer 2013, the transport pattern differed from the classic westerly transport pattern, in which air aloft at approximately 500-1500 m above ground level (AGL) flows from the Ohio River Valley, a source region for O_3 and O_3 precursors, such as NO_x. The summer 2013 transport pattern featured stronger than usual southerly flow along and east of I-95, which brought clean maritime air into Delaware and the Mid-Atlantic region, as shown in Figure 12. Winds in the transport layer aloft were shifted southerly compared to normal, with the result that large scale westerly transport from the Ohio River Valley and the Midwest was limited. The shift to southerly, rather than westerly, transport was evident at the rural monitor at Shenandoah National Park (SNP). SNP is a "sentinel" monitor for Delaware; because it is located at relatively high elevation (1000 m AGL) and west of the I-95 Corridor, SNP observes polluted air flowing into the Mid-Atlantic region from the west prior to and during high O_3 events. In 2013, SNP observed only one day, June 5, with 8-hour O_3 in excess of 60 ppbv (Figure 13), which is the lowest number of observations in recent history.

Skill of Ozone Forecasts in 2013

The skill of all O_3 forecasts in 2013 was roughly similar to recent years, although absolute errors were slightly higher due to the difficulty of forecasting for the unusually clean O_3 season. A time series of forecasts and observations for 2013 is shown in Figure 14 and error statistics for forecasts during recent years (2011-2013) are given in Figure 15. Median absolute forecast error during the 2013 season was 6.5 ppbv, which is slightly higher than recent years (5.0 ppbv in 2011 and 2012). Figure 15 also shows that there was a bias toward over-prediction of O_3 in 2013 (2.5 ppbv), which is on par with the over-prediction bias that occurred in 2011 (2.6 ppbv).

For the Code Orange O_3 cases, however, the 2013 results were below average in skill (Figures 16 and 17). Details on the calculation of skill scores are given in Appendix A. The false alarm rate was 1.0 in 2013, which is much higher than in recent years. Code Yellow (Moderate) O_3 was observed on all 5 of the days that Code Orange forecasts were issued. The total number of false alarms in 2013 (5) was less than in 2011 (10) and 2012 (6), however. The hit rate, the ratio of Code Orange forecasts to observed Code Orange, was zero, meaning that neither of the 2 observed Code Orange days was correctly forecasted with health alerts issued to the public. The very low occurrence of observed Code Orange O_3 days in 2013 is primarily responsible for the wild swing in skill compared to previous years. There were 14 observed Code Orange days in 2011 and 20 in 2012, compared to only 2 in 2013.

The high false alarm rate and concurrent low hit rate were the most striking forecast features of the 2013 summer season. While regional decreases in O_3 precursor emissions have resulted in fewer poor air quality days in Delaware since 2003, they have also increased the challenge of correctly forecasting Code Orange cases, particularly for marginal Code Orange cases (i.e., 76-79 ppbv). Prior to 2003, forecasts of peak temperature, which are quite accurate, were sufficient to identify most Code Orange days, as shown in Figure 9. However, temperature alone is no longer a necessary *and* sufficient predictor of Code Orange O_3. The "decoupling" of O_3 and temperature means that other conditions, such as local scale wind circulation, convection, cloud cover, and local emissions, are becoming greater contributing factors for observed O_3

concentrations. These phenomena, which vary on scales that numerical air quality forecast models cannot directly resolve, are more difficult to forecast, which in turn increases the difficulty of forecasting Code Orange O_3 days. To that end, 4 of the 5 false alarm forecasts in Delaware in 2013 were for marginal Code Orange levels, 76-79 ppbv. One of the two observed Code Orange days (July 19) was also at a marginal Code Orange level, 78 ppbv; the issued forecast for this day was 75 ppbv. As regional O_3 levels continue to fall in response to NO_x controls and the overall frequency of high O_3 days decreases, the likelihood of marginal observed Code Orange days will likely increase, making accurate forecasts more difficult. To help address this problem, improvements in forecast guidance are planned for 2014, as described in the "Outlook for 2014" section.

Performance of Ozone Numerical Forecast Models in 2013

As noted above and in previous reports, reductions in emissions of O_3 precursors have significantly altered the climatology of O_3 as well as the frequency of poor air quality days. These changes have reduced the efficacy of forecast methods based on long term statistical relationships, which had been used quite successfully prior to 2003. In the place of statistical guidance, forecasters have increasingly relied upon numerical air quality forecast models. These models became operational in the mid-2000s and have steadily improved in skill since that time. For the 2011-2012 O_3 seasons, a number of numerical air quality forecast models were routinely available to forecasters. The most skillful models include the NOAA-EPA air quality model (NOAA), the SUNY-Albany model (SUNY), and two versions of the Baron Advanced Meteorological Services model (BAMS) − the CMAQ and RT. In 2013, the SUNY model was not operational, and the NOAA model came close to cancellation due to federal budget reductions. As a result, the NOAA model was not updated in 2013 for emissions and other components, as had been the case in previous years. The SUNY model, which was relatively accurate in 2011-2012, was not available in 2013, so the North Carolina Department of Environment and Natural Resources (NCDENR) O_3 model was utilized instead.

Error statistics for the O_3 forecast models used as guidance in 2013 are shown in Figure 18, and skill scores for the Code Orange O_3 predictions are given in Figures 19-20. All of the forecast models suffered from an over-prediction bias ranging from 2.8-6.7 ppbv. The mean and median absolute error of the models ranged from approximately 5-10 ppbv. The most accurate models were the NOAA and NCDENR. The tendency to over-predict peak O_3 resulted in a false alarm rate for Code Orange O_3 cases, ranging from 0.67 to 0.88 (Figure 19). The high false alarm rate did improve the hit rate for the models, particularly the two BAMS variations and the NCDENR model, which were able to predict both of the observed Code Orange O_3 days. The NOAA model had the lowest false alarm rate for Code Orange O_3 forecasts in Delaware. This was a surprising result, since the NOAA model was not updated in 2013, so its performance was expected to be poor and consequently, expert forecasts were not adjusted to rely on what turned out to be relatively high quality guidance.

In weather forecasting, there is a large body of research showing that the mean, or median, forecast from a number of models, termed an ensemble forecast, is more skillful than any single model. In 2011, a pilot study was undertaken to determine if a crude ensemble of air quality forecast models could add skill to expert forecasts. The results from 2011 were encouraging enough to justify a more detailed effort in 2012. During the summer of 2012, forecasts from the NOAA, SUNY and BAMS models were combined to produce an ensemble forecast. Results showed that the ensemble forecast was more skillful than any of the individual models by all skill measures. In 2013, the missing SUNY model was replaced with the NCDENR forecast model. Because this model had not been previously used in operations on a consistent basis, its impact on the ensemble forecast skill was unknown. As it turned out, the 2013 ensemble overall skill in Code Orange O_3 cases was better than any of the component models, with a false alarm rate of 0.60 and a hit rate of 1.00 . In 2014, an adjusted ensemble will be tested, which will weight the models based initially on their forecast skill from 2013.

Outlook for 2014

Two major improvements in forecast guidance are planned for 2014. First, the ensemble model guidance will be improved by the full inclusion of the NCDENR forecast model and by moving from a straight mean ensemble value to a weighted average of the three models. This weight will be initially based on 2013 performance and then adjusted as the 2014 results accumulate. In addition, funding was reinstated for the NOAA forecast model, so the emissions should be updated in the model for 2014, which will improve model performance. There has been no word on a resumption of the SUNY forecast model, so the NCDENR model will continue to be used in the ensemble. Second, a new statistical model, trained on the period from 2008-2012, will be tested in 2014. The data for the training model have been collected and organized into a single dataset, and tests are underway to create the most accurate statistical model. These forecast guidance improvements are expected to increase the accuracy of all expert O_3 forecasts in 2014, and hopefully they will make Code Orange O_3 days easier to identify, particularly marginal Code Orange O_3 cases.

The lingering issue related to the dissemination of Air Quality Alerts (AQAs) through NWS forecast offices for Code Orange air quality days appears to have been resolved. With assistance from staff at NOAA headquarters, the Mid-Atlantic air quality forecasters were able to prevail upon the Mount Holly and Sterling local NWS forecast offices to post AQAs for Code Orange forecasts on the main NWS “Watches and Warnings” page. This change occurred during the mid-summer of 2013 and will be fully in effect in 2014.

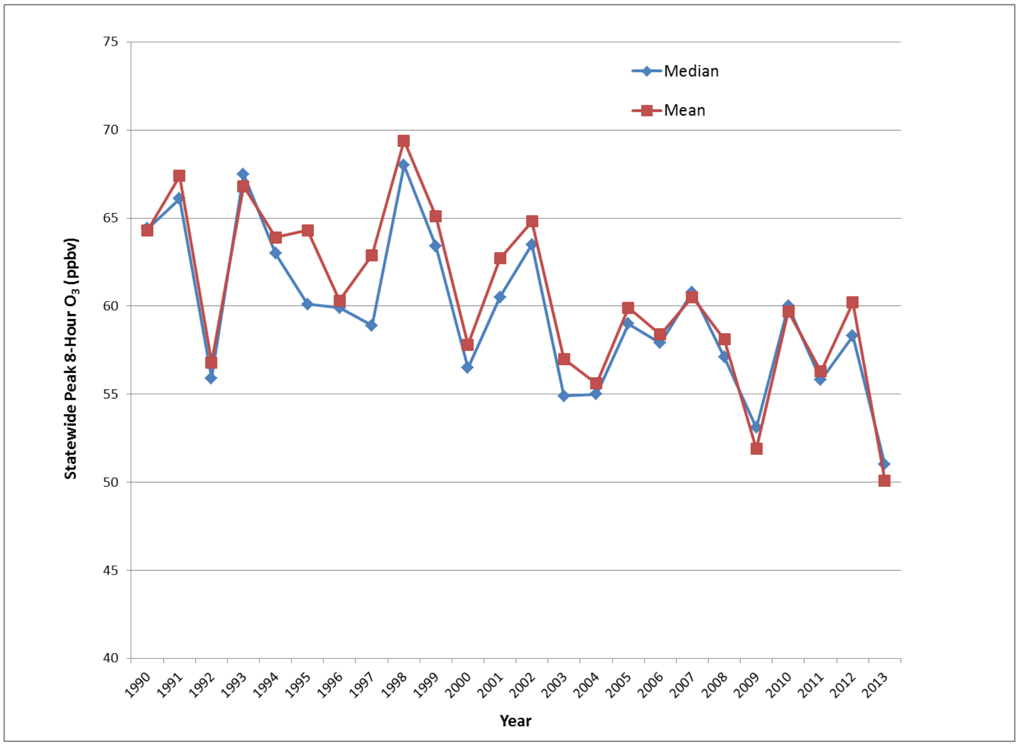

Figure 1. Seasonal (May-September) mean and median peak 8-hour observed O_3 in Delaware for 1990-2013.

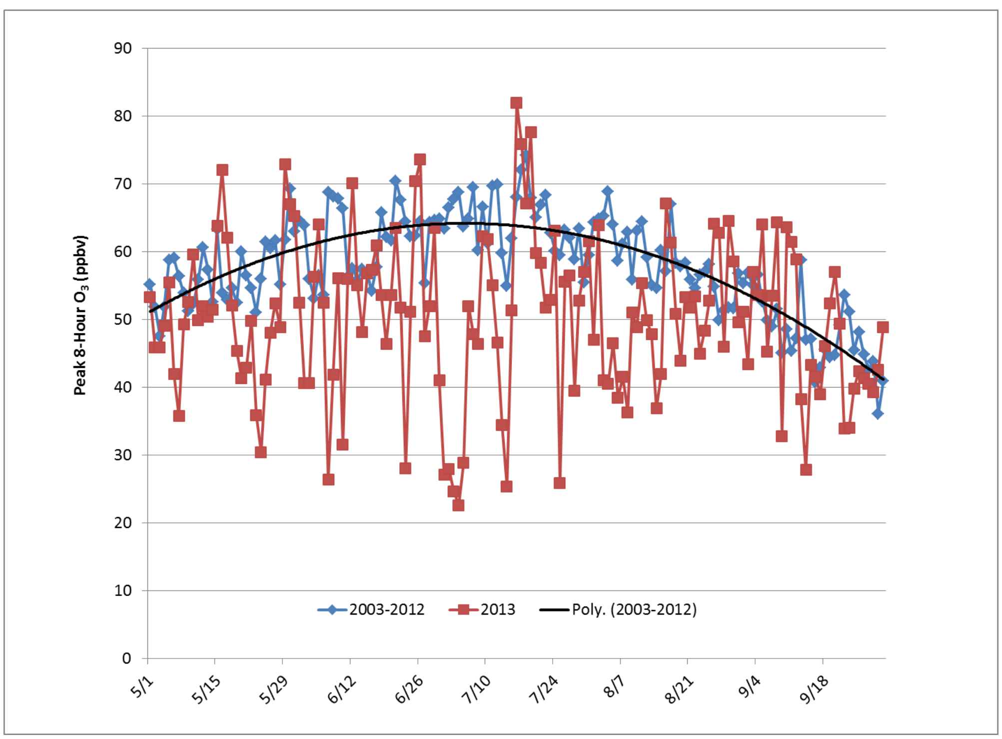

Figure 2. Daily time series of peak 8-hour observed O_3 in Delaware for 2013 (red line) compared to the 2003-2012 average (blue line). The black line is the best polynomial fit to the 2003-2012 average.

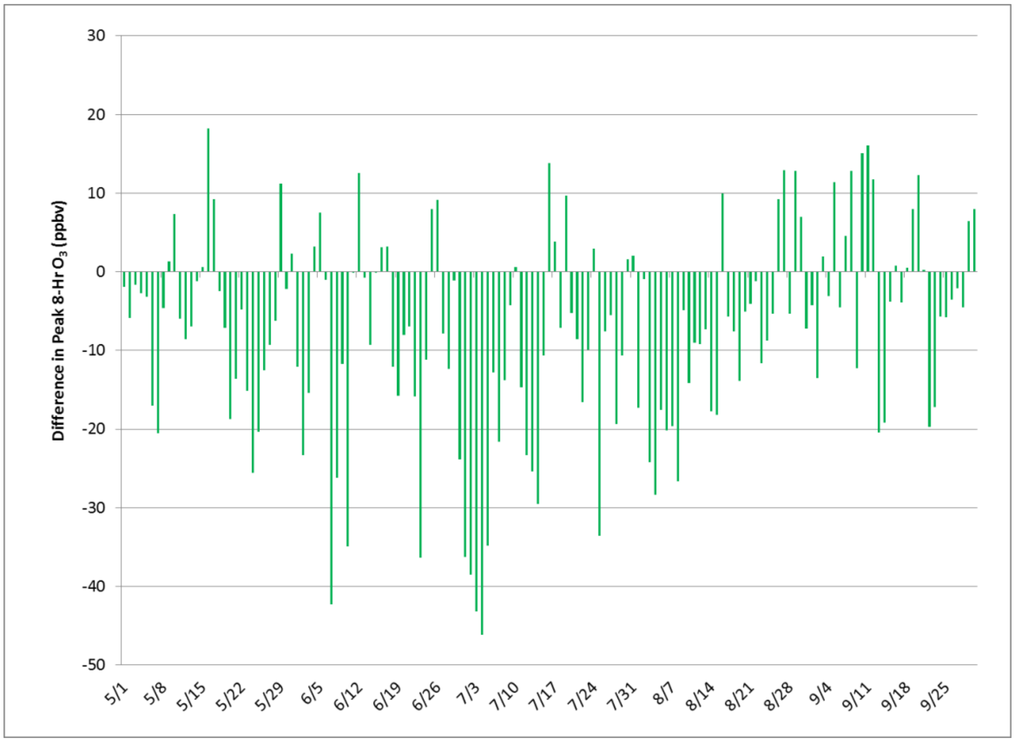

Figure 3. Daily time series of the difference between peak 8-hour observed O_3 for 2013 compared to the 2003-2012 average.

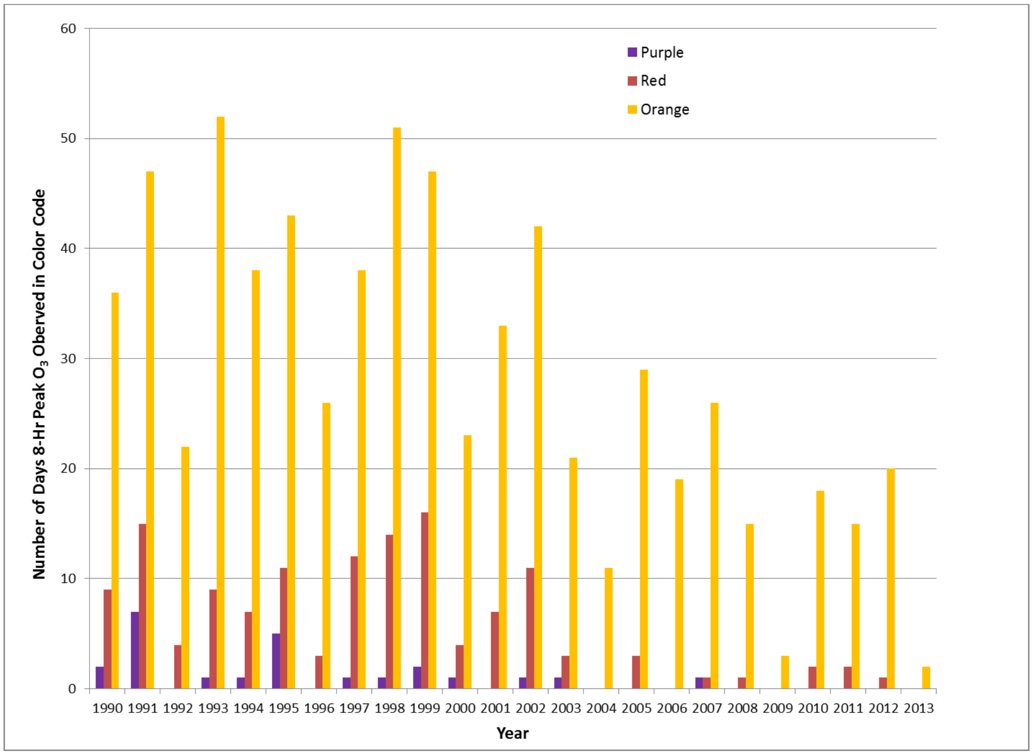

Figure 4. Frequency of Code Purple, Code Red, and Code Orange observed O_3 days in Delaware for 1990-2013.

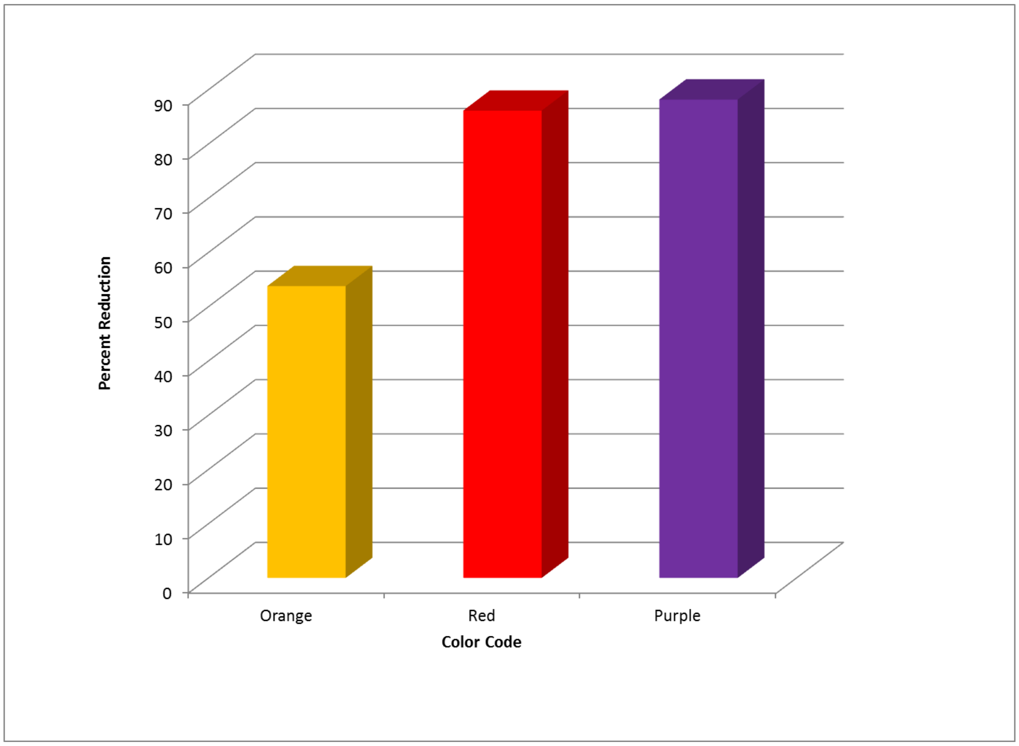

Figure 5. Percent reduction in the occurrence of Code Orange, Red, and Purple observed O_3 days in Delaware since 2002.

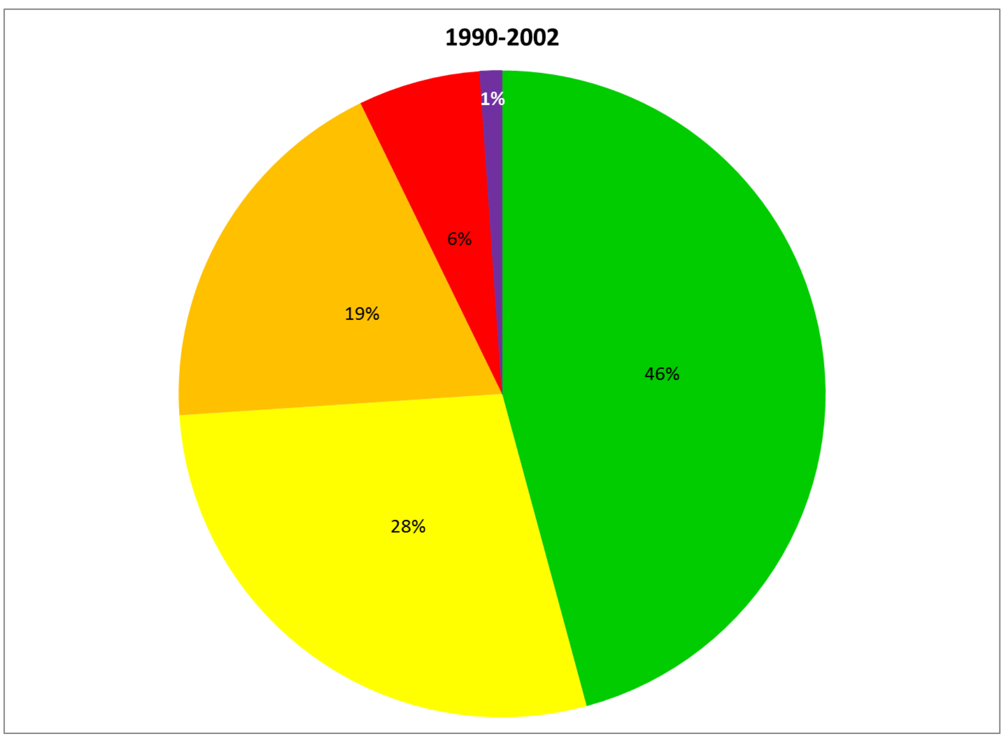

Figure 6. Frequency of AQI color codes for observed O_3 in Delaware for 1990-2002.

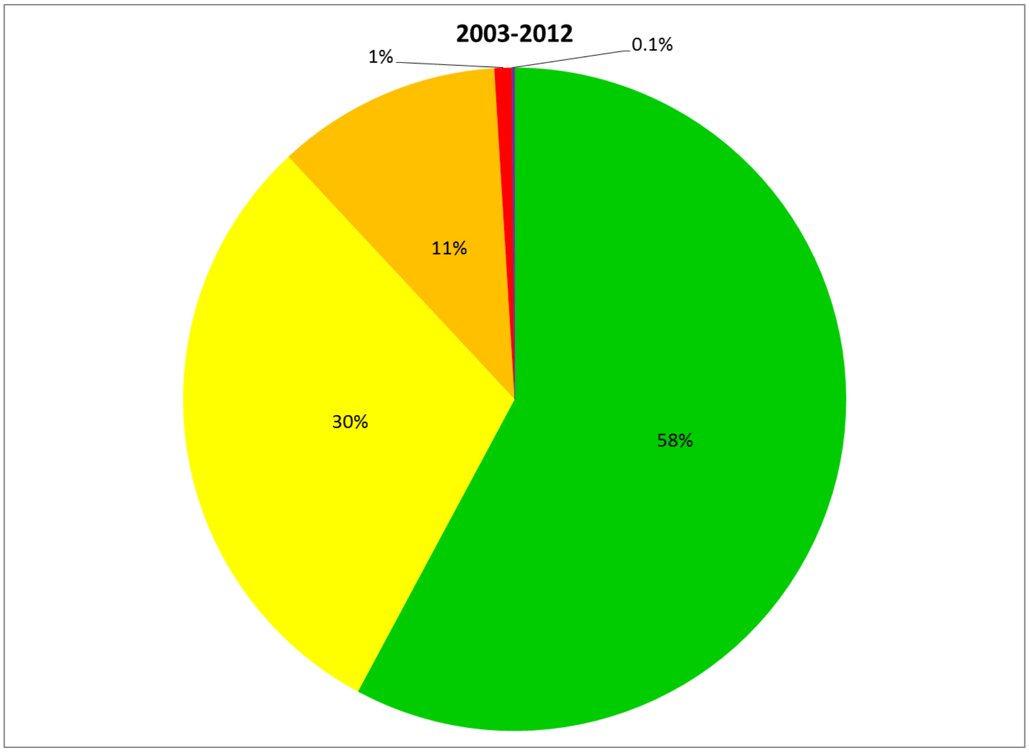

Figure 7. Frequency of AQI color codes for observed O_3 in Delaware for 2003-2012.

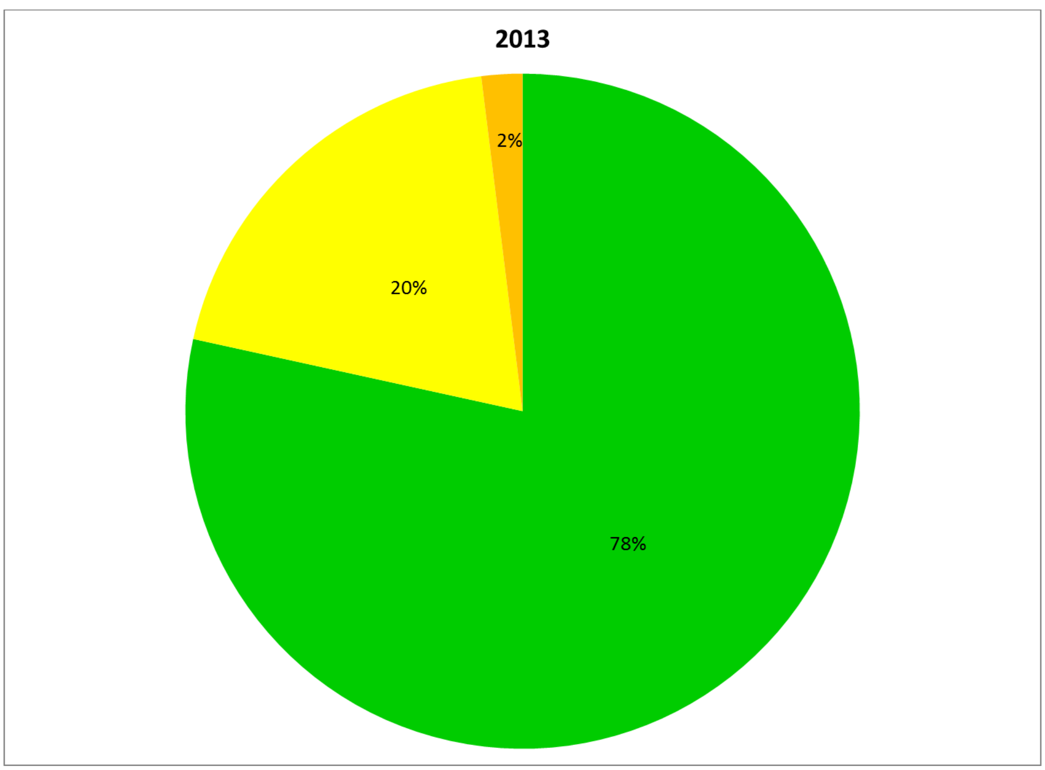

Figure 8. Frequency of AQI color codes for observed O_3 in Delaware for 2013.

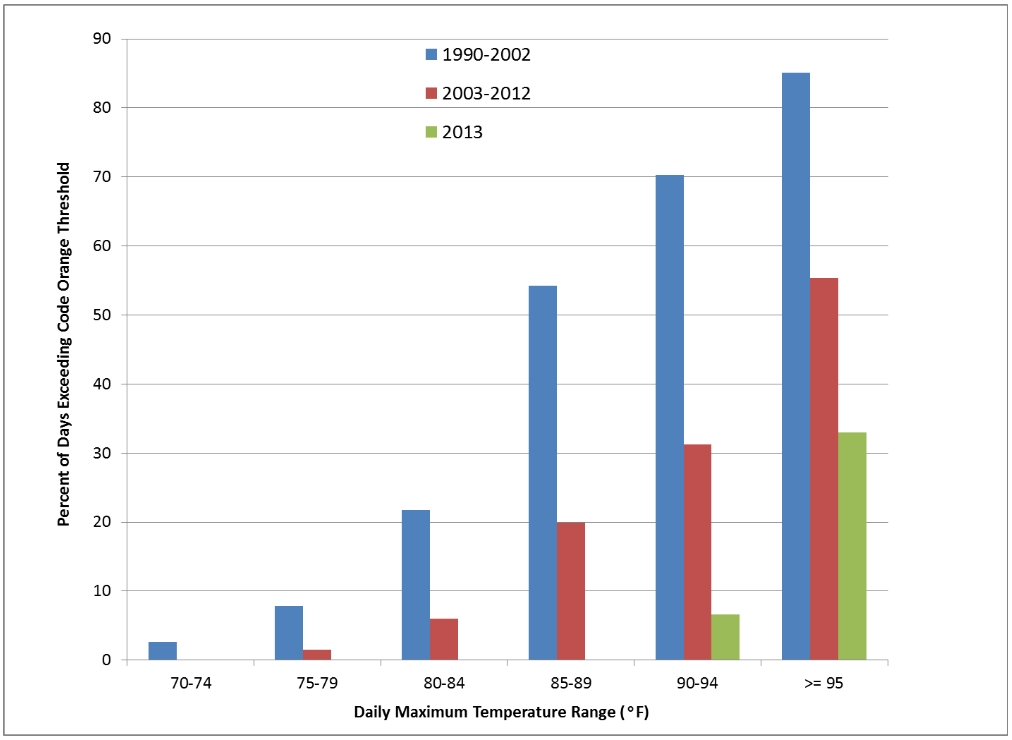

Figure 9. Percent of days exceeding the observed O_3 Code Orange threshold in Delaware for bins of maximum surface temperature measured at Dover, Delaware (DOV).

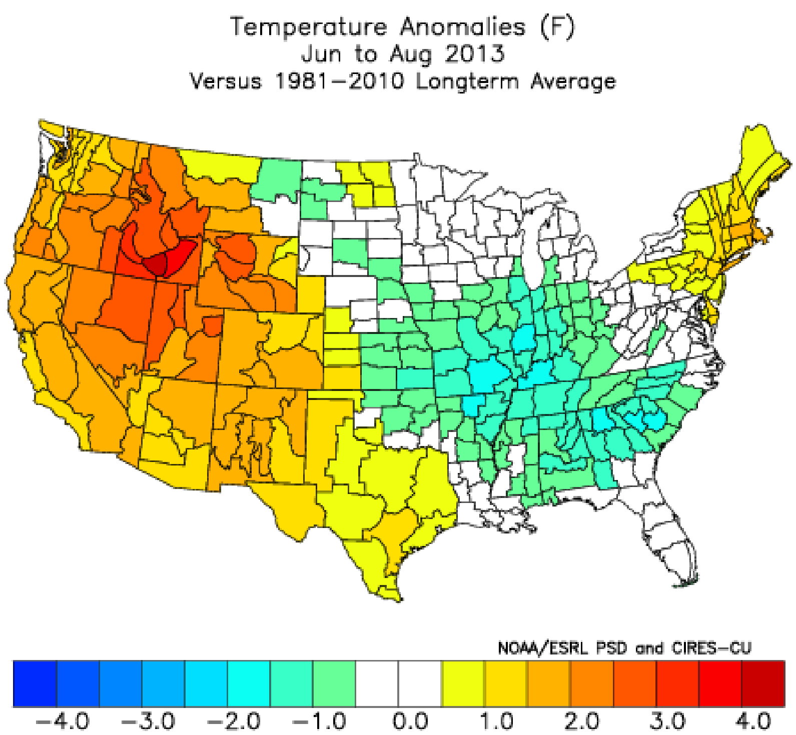

Figure 10. Temperature anomalies (in °F) in the U.S. for June-August 2013 compared to the 1981-2010 average (courtesy of NOAA/ESRL).

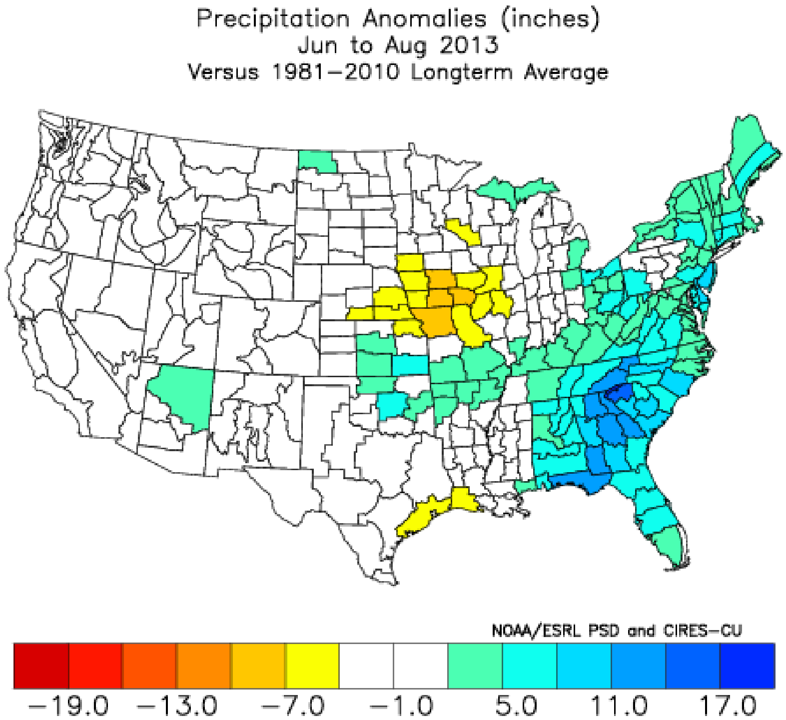

Figure 11. Precipitation anomalies (in inches) in the U.S. for June-August 2013 compared to the 1981-2010 average (courtesy of NOAA/ESRL).

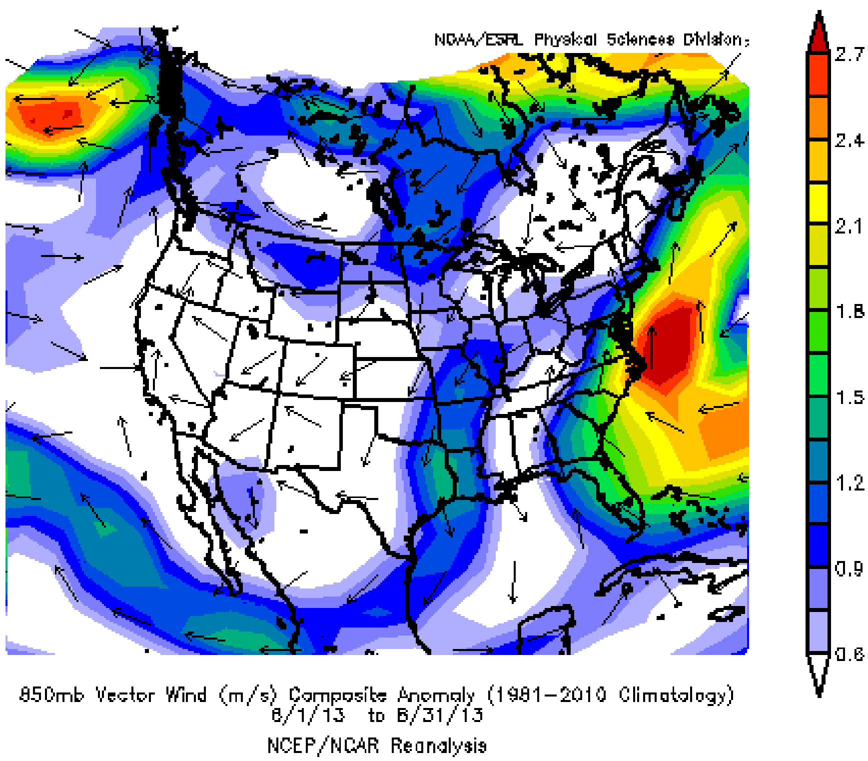

Figure 12. Vector wind anomalies at 850 mb (~1500 m AGL) in the U.S. for June-August 2013 (courtesy NOAA/ESRL).

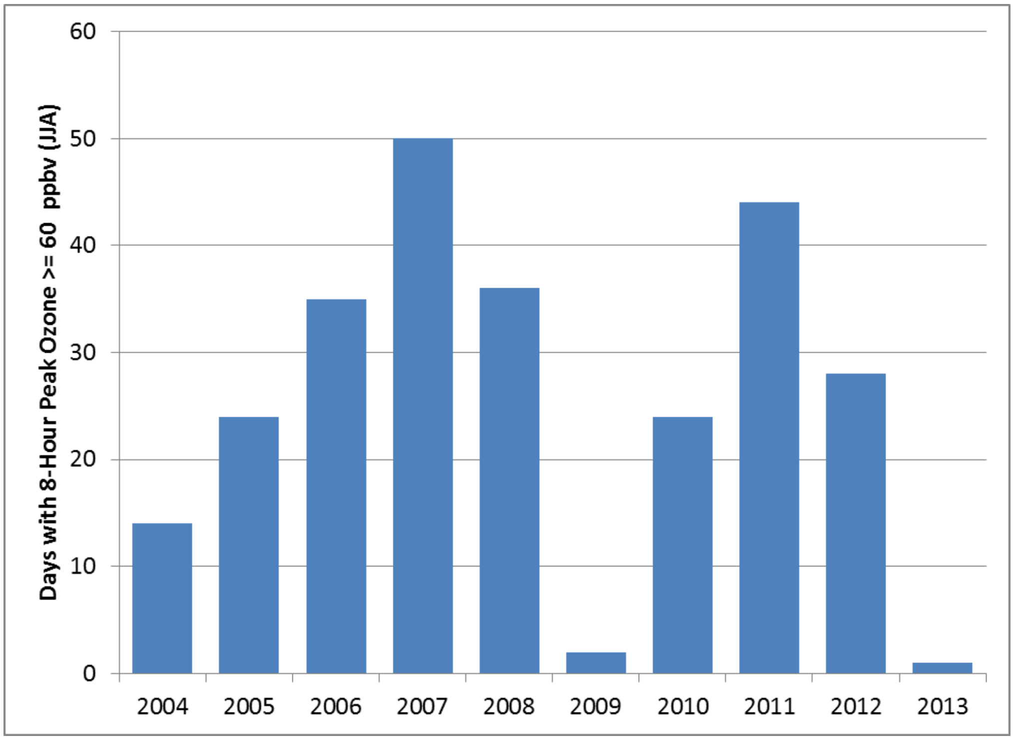

Figure 13. Number of days with peak 8-hour observed O_3 above the 60 ppbv threshold at Shenandoah National Park, VA for June-August of 2004-2013.

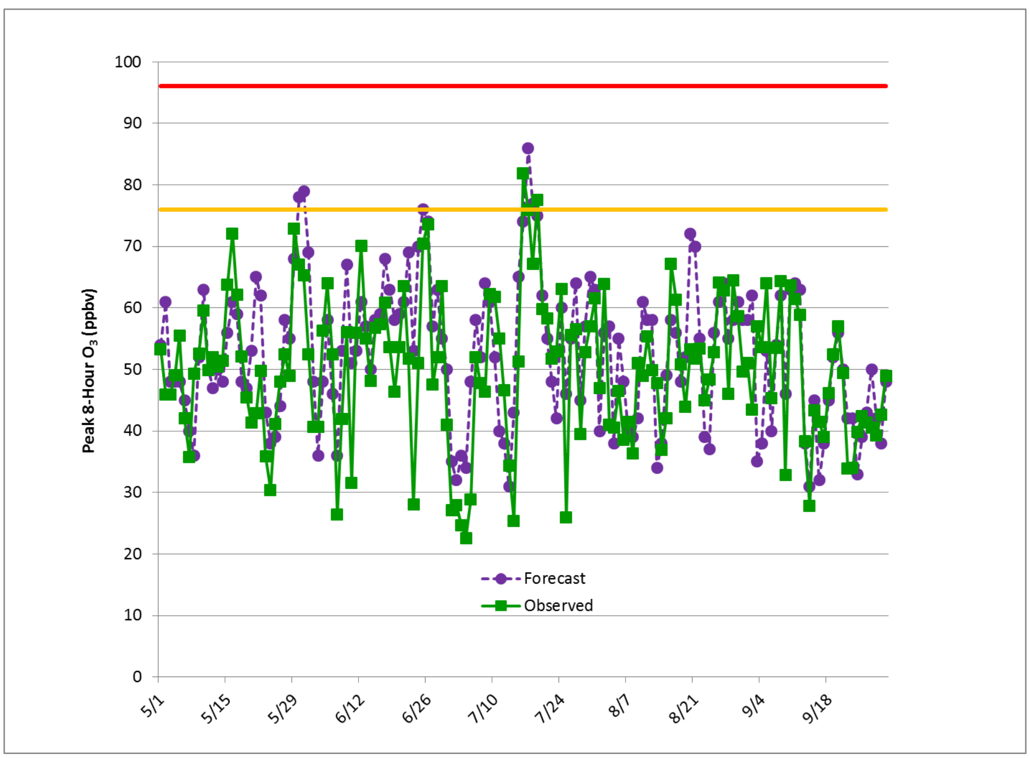

Figure 14. Peak 8-hour O_3 forecasts and observations for Delaware during May-September, 2013. The orange and red lines indicate the Code Orange and Code Red thresholds, respectively.

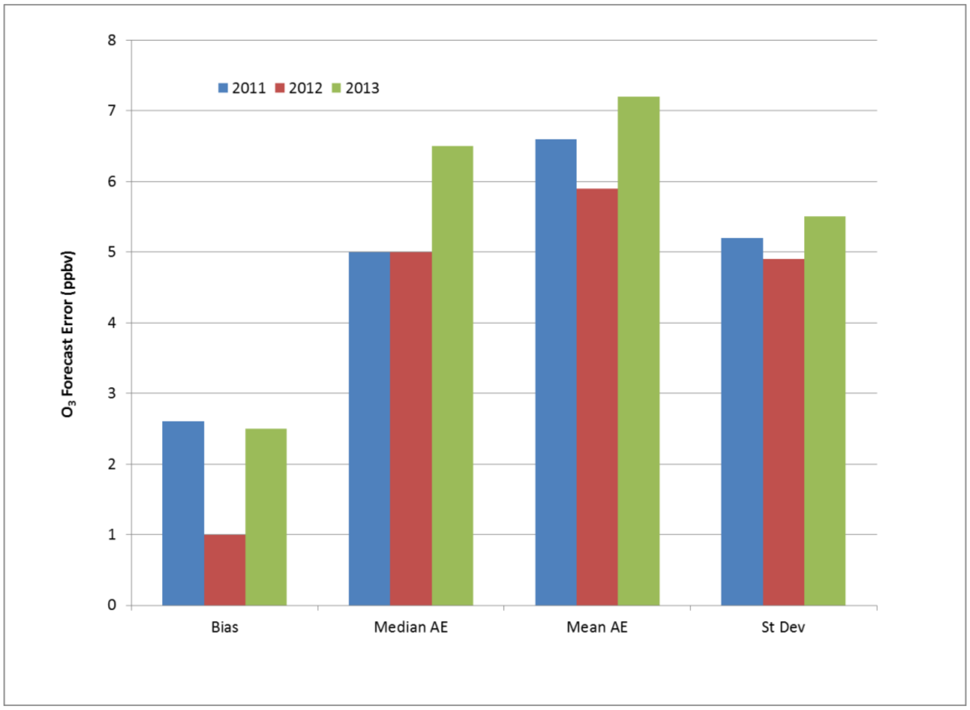

Figure 15. Error statistics for all peak 8-hour O_3 forecasts in Delaware for 2011-2013. "Median AE" refers to median absolute forecast error, "Mean AE" refers to mean absolute error, and "StDev" refers to the standard deviation of the mean absolute error.

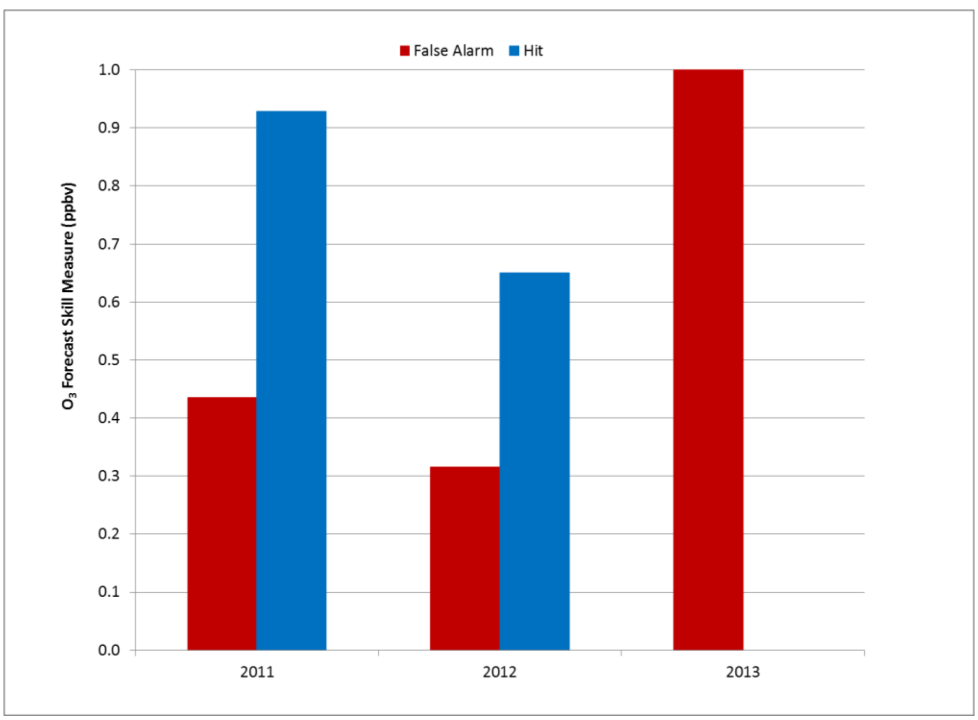

Figure 16. False alarm rate and hit rate for Code Orange O_3 forecasts in Delaware for 2011-2013.

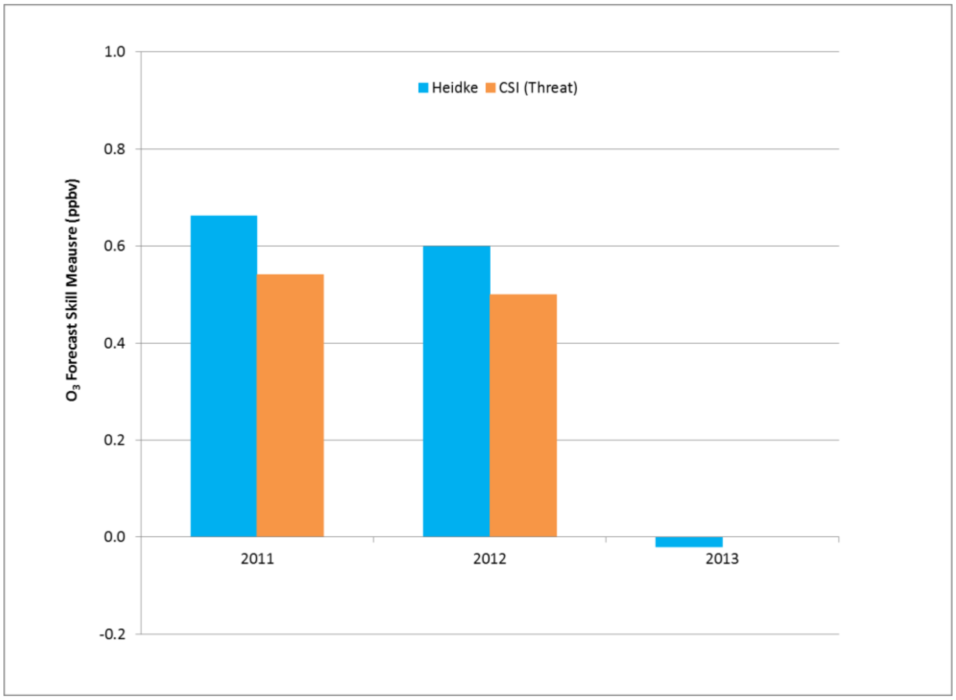

Figure 17. Heidke skill scores and CSI (threat scores) for Code Orange O_3 forecasts in Delaware for 2011-2013.

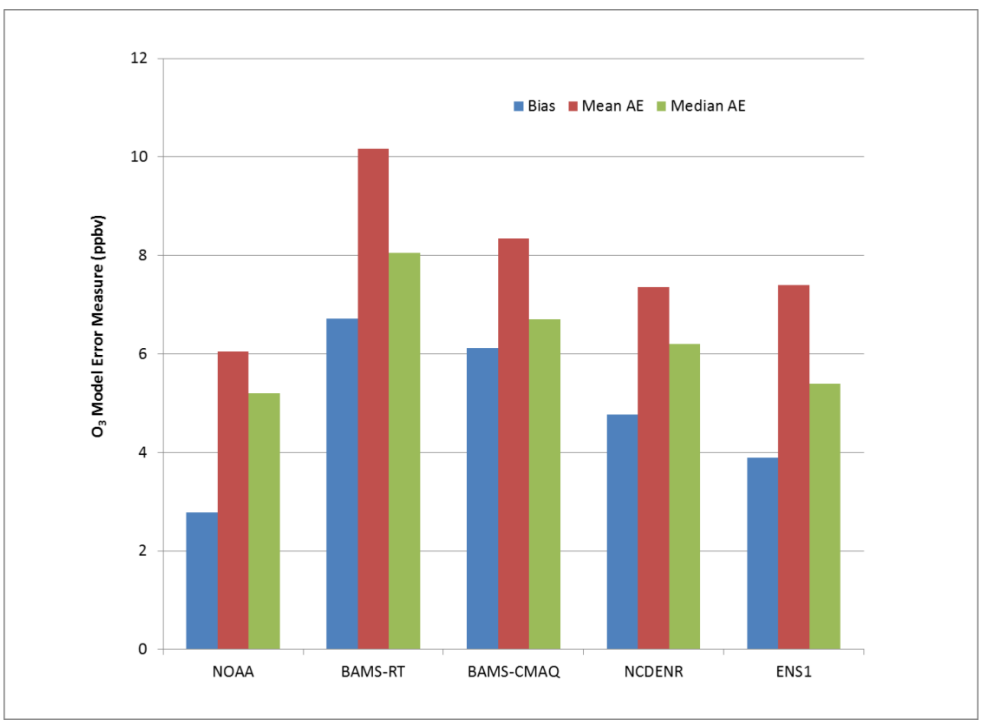

Figure 18. Error statistics (as in Figure 15) for air quality numerical forecast model guidance for all peak 8-hour O_3 predictions in Delaware in 2013. Two variations of the Baron Meteorological Services (BAMS) models are shown (CMAQ and RT). The ENS1 is a mean value from the NOAA, BAMS-RT, BAMS-CMAQ, and NCDENR model forecasts.

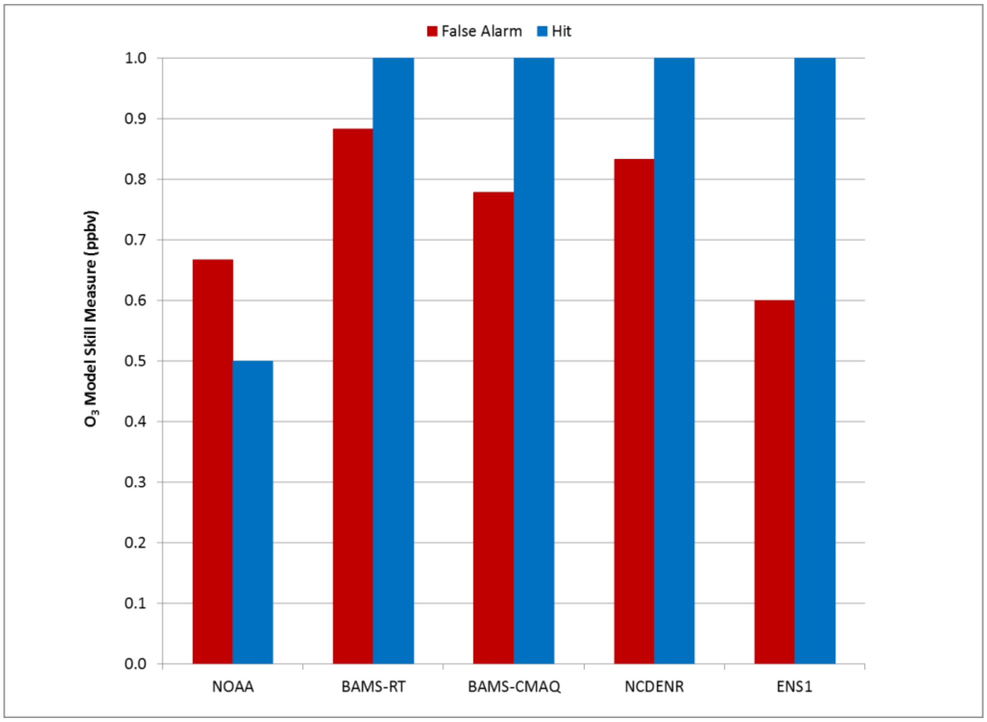

Figure 19. False alarm rate and hit rate for Code Orange O_3 predictions by air quality numerical forecast models in Delaware in 2013.

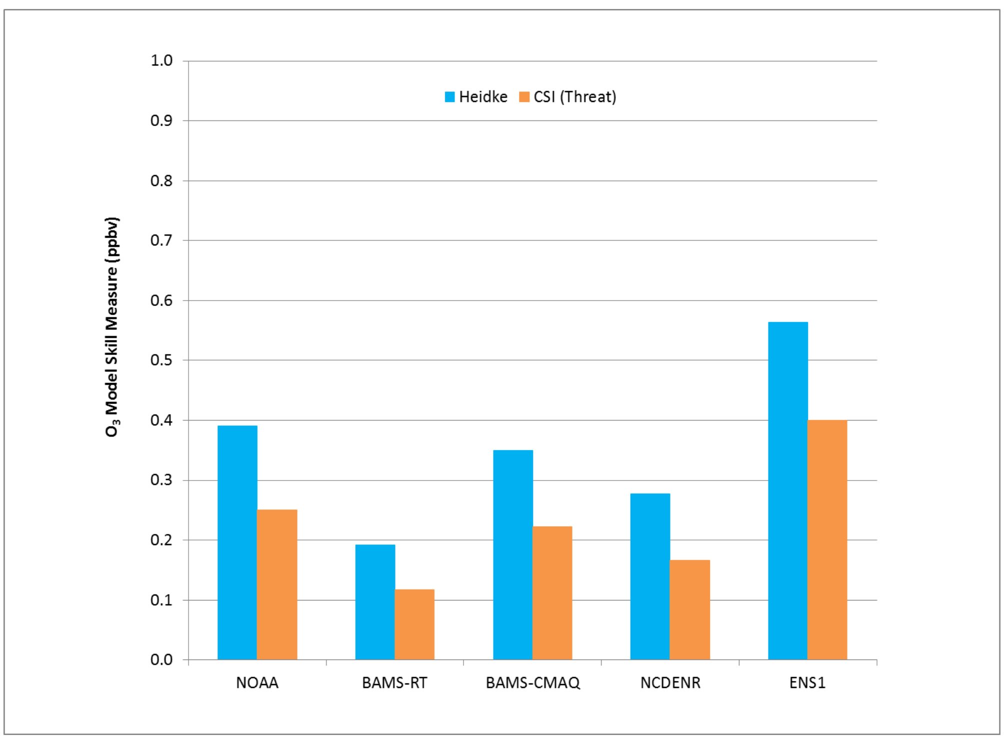

Figure 20. Heidke skill scores and CSI (threat scores) for Code Orange O_3 predictions by air quality numerical forecast models in Delaware in 2013.

Appendix A. Skill Measures for Threshold Forecasts

The determination of the skill of a threshold forecast (e.g., Code Orange air quality) begins with the creation of a contingency table of the form:

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

For example, if Code Orange O_3 concentrations are both observed and forecast ("hit"), then one unit is added to "a." If Code Orange O_3 is forecast but not observed ("false alarm"), then one unit is added to "b."

Basic Skill Measures

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

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

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.

False Alarm Rate (F) = $\frac{b}{a+b}$

This is a measure of the rate at which false alarms (high O_3 forecast but not observed) occur.

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

The hit rate is often called the "probability of detection"

Miss Rate = 1 – H

Correct null forecasts:

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

Accuracy:

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

Other Skill 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).

$$\textbf{Skill Score (SS}_{\textbf{ref}}\textbf{)} = \left(\frac{A - A_{ref}}{A_{perf} - A_{ref}}\right) * 100\% = \text{ nn}\%$$

The skill score is typically reported as a percent improvement of accuracy (A) with respect to a reference forecast. The reference forecast accuracy (A_{ref}) is typically climatology or persistence. The perfect forecast (A_{perf}) is usually 1 (e.g., for hits) or 0 (e.g., for false alarm).

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.

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

For this measure, the range is [-1,1] with a random forecast equal to zero.

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

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

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:

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

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, 467 pp., 1995.