Memo

To: Air Quality Index Forecasters, IBEAM Programmers

From: Wayne L. Cornelius

CC: DAQ Monitoring Staff, Jeff Francis, Patrick Reagan

Date: 2012-09-18

Re: **Ozone Surrogate Data Imputation for the Air Quality Index in North Carolina**

Summary

This note provides criteria for computing "surrogate ozone data" in support of an estimate of the Air Quality Index for the time period beginning five hours before the hour of the report and ending three hours after the hour of the report. The method is defined and then applied to ozone monitors operated in North Carolina and acquired by the NC Division of Air Quality during 2011.

Contents

List of Figures

List of Tables

Introduction

This note provides criteria for computing "surrogate ozone data" in support of an estimate of the Air Quality Index (AQI) for the current time, that is, an imputed 8-hour average covering the period beginning five hours before the hour of the report and ending three hours after the hour of the report. The method is defined and then applied to ozone monitors operated in North Carolina and acquired by the NC Division of Air Quality (DAQ) during 2011. A map of site locations is shown in Figure 1.

The ozone subindex of the AQI is based on a step function relating the 8-hour average concentration (in parts per billion, or "ppb") to a dimensionless number that signifies "good" air quality in the range from 0 to 50, "moderate" air quality in the range from 51 to 100, and progressively worse categories of air quality for index values from 101 to 500. The exact relationship between an ozone concentration and its index number is not a specific concern in this note.

The nature of data acquisition dictates that a *complete* 8-hour pollutant concentration cannot be computed until 8 hours and a few minutes have elapsed after the starting time of the average. This means that an actual ozone concentration or AQI index value is past or expired by the time it is reported. It is perceived to have been "current" fully five hours earlier, the midpoint of the time interval covered by the reported average[1]. Such averages are nonetheless useful, because they represent air pollutant concentrations measured within the past 24 hours and thus contribute to the AQI reported for that period and previously forecasted by DAQ staff.

However, AQI forecasters and customers are often interested in an AQI that is accurate for a time period close to covering from four or five hours in the past up to three or four hours in the future. This need is typically met by application of a surrogate 8-hour average based on the current *one*-hour average (that is, the average received between minute 0 and minute 59 of the hour immediately[2] preceding the report time).

---

[1] EPA guidance on computing pollutant concentrations usually allows averages that are "75 percent complete" to be regarded as valid. This means an ozone average from seven hours in the past or from six hours in the past to the present moment is considered an acceptable 8-hour average as it stands, even though one or two hours of these averages are not yet known. An average with only five known hours and three unknown hours fails the 75% criterion.

[2] AIRNow Frequent Questions, "How are your ozone maps calculated?", published online at http://airnow.supportportal.com/link/portal/23002/23002/Article/16115/How-are-your-ozone-maps-calculated, accessed 25 July 2012.

Ozone concentrations rise and fall in diurnal cycles, correlated with temperature, sunlight, automotive emissions and other factors. Because of this, the amplitude of 8-hour averages is smaller than the amplitude of 1-hour averages, although the long-run expected values of 1-hour and 8-hour averages are equal. (Even a 24-hour average of overlapping 8-hour averages will be nearly equal to the average of the 1-hour averages for the same 24 hours.)

It follows that an *instantaneous* ozone 8-hour surrogate concentration should be exactly the same as the current 1-hour average. However, when the 1-hour concentration is at a daily peak value, the surrogate current 8-hour concentration should be less than the simultaneous 1-hour concentration, and when the 1-hour concentration is at a daily minimum value, the surrogate current 8-hour average should be greater than the simultaneous 1-hour concentration.

There is more interest in current AQI values when the current concentration is high. Therefore, the surrogate concentration estimator is derived by using the daily maximum 1-hour average to predict the daily maximum 8-hour average and applying this prediction equation to the current 1-hour average at any time of the day. A consequence of this practice is that the actual 8-hour concentration observable 4 hours later will usually be less than the point-value concentration predicted by the surrogate equation. However, a technical assistance recommendation mitigates this by advising forecasters not to predict the level of the AQI, but only the *category* represented by the daily maximum 8-hour average using a confidence interval centered on the current one-hour average, along with the forecaster's determination of whether the AQI is expected to rise or fall from the current predicted value during the hours being predicted[3].

Methods

I acquired daily maximum ozone averages, both 1-hour and 8-hour, by running the IBEAM Report, Ambient Monitoring: Ozone 8 Hour Average for a Time Period, One or All Sites. This report produces maximum 8-hour averages and maximum 1-hour averages, all properly adjusted for validity flags and required completeness. I attempted to query all sites, for the ozone season April 1 through October 31, 2011. However, I was unable to export the dataset thus queried, apparently because of its size, so I queried 40 ozone sites one-at-a-time, imported the files with S+® and combined them into a single S+ data object.

Using S+, I applied four regression models to the data, and tabulated coefficient estimates and their standard errors and the reported significance test probability for the null hypothesis that the intercept of the linear equation equals zero. Two of the models are applied to the full statewide database as a whole, and two produce regression estimates for each individual site.

Model 1. E(Max.8.Hr.Avg) = **RATIO** * Max.1.Hr.Avg

[3] EPA OAQPS (1999). Guideline for Reporting of Daily Air Quality – Air Quality Index (AQI). EPA-454/R-99-010. Mintz, David (2009).Technical Assistance Document for the Reporting of Daily Air Quality – the Air Quality Index (AQI). EPA-454/B-09-001, URL: http://www.epa.gov/airnow/aqi_tech_assistance.pdf, accessed 10 Sep 2012.

Model 2. E(Max.8.Hr.Avg) = **INTERCEPT** + **SLOPE** *Max.1.Hr.Avg

Model 3. E(Max.8.Hr.Avg) = **RATIO(SITE)** * Max.1.Hr.Avg

Model 4. E(Max.8.Hr.Avg) = **INTERCEPT(SITE)** + **SLOPE(SITE)** *Max.1.Hr.Avg

AIRNow Tech has used Model 4 to generate most ozone surrogate averages and Model 1 in those cases where the dataset of averages was deemed inadequate[4]. AIRNow Tech analyzed their entire national database of ozone data in 1994-96 and 2008-10, fitting regressions to three consecutive years of data, if available, and thereafter applying the exact Model 4 equation to generate ozone surrogate averages. For monitors that did not have three years of data available to analyze, AIRNow Tech applies Model 1 with **RATIO** set equal to 0.85.

In the next section, I analyze whether the **RATIO** estimate is within 2.5 standard errors of each **RATIO(SITE)** estimate, whether each **INTERCEPT(SITE)** estimate is significantly different from 0.0, and whether the **SLOPE** estimate is within 2.5 standard errors of each **SLOPE(SITE)** estimate. I use these findings to recommend site-specific surrogate estimators.

Results

The dataset I analyzed contained 8,189 valid monitor-days of data. The days included are from April 1, 2011 through October 31, 2011, omitting site-days for which no valid 8-hour maximum is available.

Figure 2 shows an overall statistical summary of the daily maximum 1-hour and 8-hour averages. The arithmetic mean concentrations are 54.30 ppb 1-hour and 48.87 ppb 8-hour.

The standard error of the 8-hour averages is 12.69, while the standard error of the residuals ranges from 2.839 for Model 4 to 2.946 for Model 1. Table 1 shows regression estimators in which the daily maximum 8-hour average is the response variable, and the daily maximum 1-hour average is the predictor variable. Note the overall estimate of **RATIO** (Model 1) is 0.900205, differing from the ratio of the arithmetic means (48.869/54.304 = 0.899915) by about 0.03 percent.

I highlighted certain entries in Table 1 and locations in Figure 1 based on the appearance that the site regression coefficients are significantly different from the overall regression estimates. All four of the highlighted ratio statistics are from sites in the mountain regions of NC. The determination of significance was more casual than rigorous: I calculated the absolute difference between the site ratio estimate and the overall ratio estimate, and the absolute difference between the site slope estimate and the overall slope estimate and highlighted those cases where the difference exceeded 2.5 standard errors of the site estimate. A difference of 2.5 standard errors corresponds to about 1.2 percent probability of erroneously declaring significance if the error has a Gaussian distribution, although I did not rigorously verify the Gaussian distributional assumption. (If the error distribution is not actually Gaussian, the

[4] Information is published online at http://www.airnowtech.org/Resources.cfm, accessed 25 July 2012.

probability of an erroneous decision might be different from 1.2 percent, but the basis for the decision about statistical significance is still the same.)

Recommendations

In all 35 cases not highlighted in Table 1, the *Overall* **RATIO** statistic is recommended for the surrogate concentration estimator, (8-hour average = 0.9 × 1-hour average) (Model 1).

For the four highlighted sites with a highlighted **RATIO(SITE)** statistic, the highlighted ratio should be used as the surrogate concentration estimator (8-hour average = **RATIO(SITE)** × 1-hour average) (Model 3). For these sites, the ratio estimates are approximately 0.93. High elevation sites tend to exhibit lower amplitudes of diurnal variation than do average and lower elevation sites. The expected value of the ratio of maxim 8-hour averages to maximum 1-hour averages is therefore closer to 1 for higher elevation sites than it is at typical lower elevation sites.

For Garinger, the intercept and slope statistics are highlighted, and the recommendation is that this linear equation (8-hour average = 2.5 ppb + 0.846 × 1-hour average) should be used for the surrogate concentrations from that site (Model 4). An inspection of the Garinger data, shown in Figure 2, shows 8-hour averages are consistently less than 0.9 x 1-hour averages when the 1-hour averages are greater than 80 ppb, while the Model 4 equation for Garinger underestimates the 8-hour average when the 1-hour average is between 50 and 80 ppb. A better regression model for Garinger would be a curved line or a segmented straight line. However, ratio or simple linear regression estimates are good for all other sites and adequate for Garinger, and it is inconvenient to adopt a more complicated model for the benefit of only this one site.

Software Implementations

DAQ has been using a version of E-DAS (registered to either ESC Corp., until 2006, or Agilaire Corp. LLC) for management of monitoring data since before 2002. In 2012, DAQ will replace E-DAS with Agilaire's AirVision product. DAQ reports real time AQI using the algorithms programmed in this software package. In this section, I describe the implementation of the AQI provided by this software. Software constantly evolves, so it is necessary to stipulate that this discussion refers to E-DAS version 5.52 and AirVision version 2.6.31. For E-DAS, version 5.52 is the terminal version, but Agilaire may change AirVision's specification of the AQI in any future version.

Both E-DAS and AirVision allow users to program a moving 8-hour average of pollutant concentrations and a step function of the type required for AQI reports, as well as appropriate methods for aggregating several such averages and determining the maximum result that should be reported to the public. AirVision is prepopulated with ozone AQI breakpoints corresponding to the *upper* limits of each descriptor category defined in the 2008 NAAQS.

Both E-DAS and AirVision allow 8-hour averages to be calculated as either forward averages (the time stamp of the average is the first hour of the 8-hour interval) or backward averages (the time stamp of the average is the 8th hour of the 8-hour interval). However, the intent of AIRNow reporting is to report AQI index values based on centered averages (8-hour averages would be time stamped at the 5th hour of the interval), and neither E-DAS nor AirVision offers this capability.

Both E-DAS and AirVision provide for the calculation of surrogate ozone concentrations using linear equation coefficients, but they differ in how they apply the equations, and the application is not necessarily in conformance with the AIRNow specification.

E-DAS appears to be able to calculate 8-hour averages with 6 or 7 valid hours, but not with fewer than 6 valid hours. When the ozone surrogate calculation is "enabled", all AQI calculations are the standard *8-hour* averages adjusted for the equation. The calculation is not applied to a one-hour average. It is not a surrogate value at all, but merely an (unnecessary) adjustment to an average that is already valid as it stands.

AirVision is set up to calculate 8 hour averages with ≥6 hour completeness (the number of required valid hours can be changed by means of a configuration setting), and will substitute the one-hour surrogate average for *all* hours with less than the required completeness.

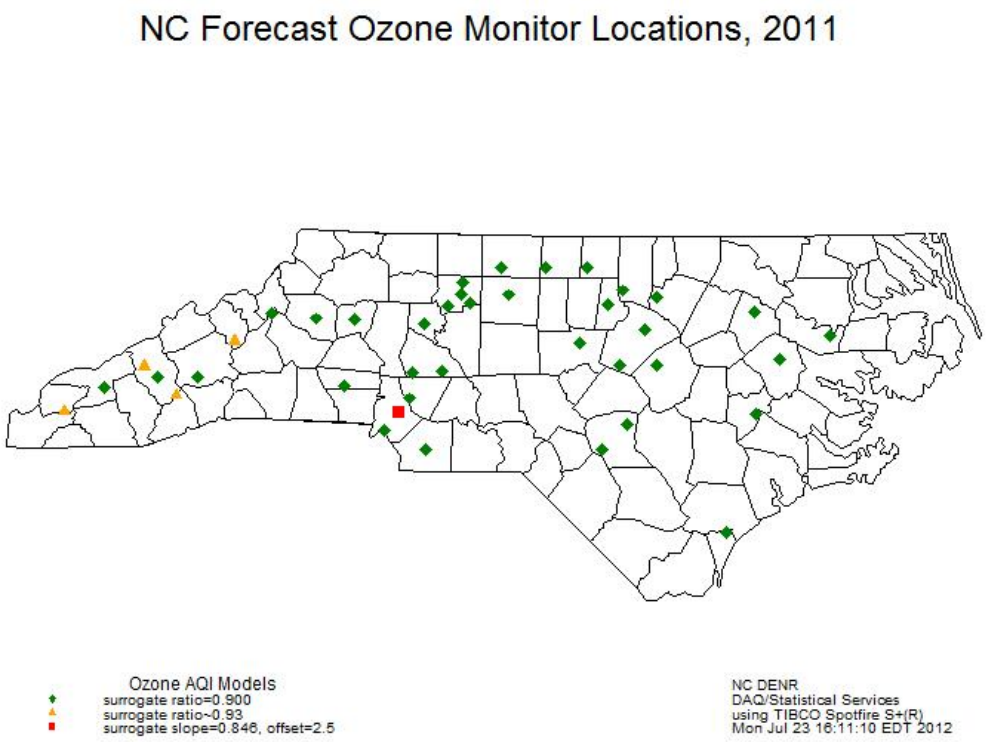

Figure 1. North Carolina Forecast Ozone Monitor Sites.

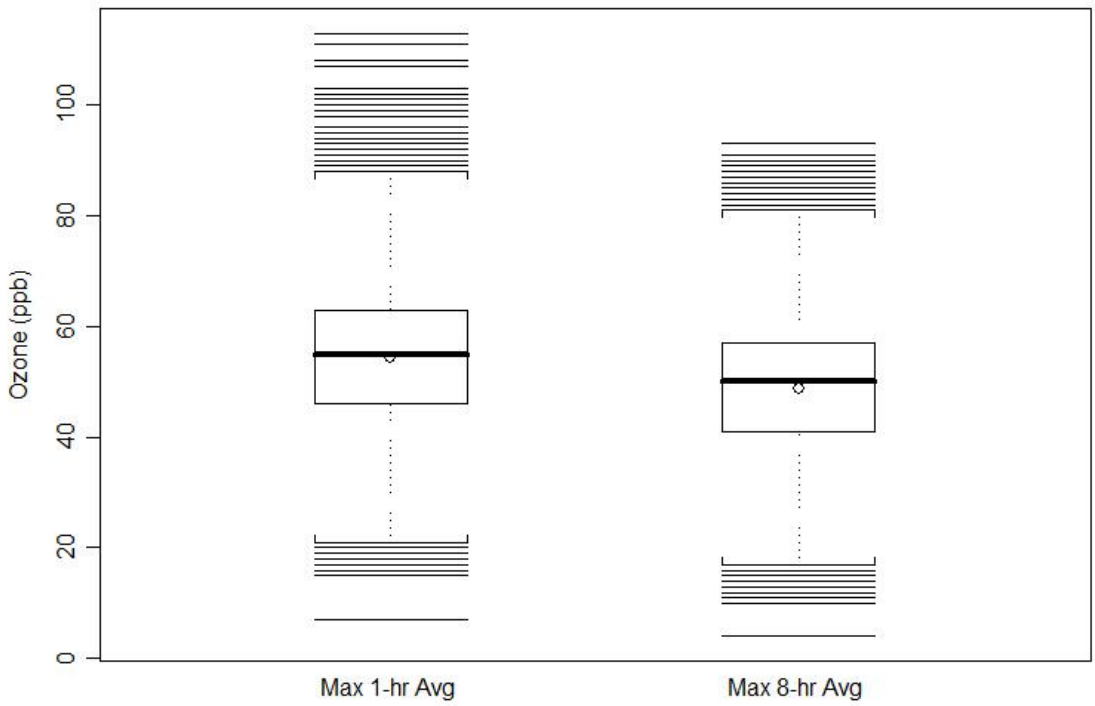

Figure 2. Statewide Distributions of Daily Maximum Ozone Concentrations, 2011.
Heavy lines mark the median concentrations (55 ppb 1-hour and 50 ppb 8-hour), and circles mark the arithmetic mean concentrations (54.3 ppb 1-hour and 48.9 ppb 8-hour). The interquartile ranges are 17 ppb 1-hour and 16 ppb 8-hour.

Table 1. Surrogate Equation Coefficients.
Ratio is the estimated slope of a linear regression forced through 8-hr max = 0 at 1-hr max=0. Intercept and Linear Slope are the regression coefficients of an unconstrained regression. Each SE statistic is the standard error of the corresponding statistic. Prob(inter.) is a test statistic for rejection of the null hypothesis that the Intercept = 0, based on the assumption that the distribution of the standard error is Gaussian. If this probability is very small, it constitutes evidence that the regression intercept value is significantly different from zero (but I do not necessarily adopt the indicated equation as the surrogate estimator—see the main text). Highlighted entries are site statistics for which the *Overall* statistic is not within 2.5 standard errors, suggesting (though not rigorously showing) that the highlighted statistic is significantly different from the corresponding *Overall* statistic. The preferred surrogate equation is the *Overall* Ratio (Model 1 in the text) for all sites that have no highlighted entries.

Site	Ratio	SE	Intercept	SE	Prob (inter)	Linear Slope	SE
Overall	0.900205	0.0006	-0.3	0.13	0.05	0.905	0.0024
Arrowood	0.884	0.0035	1.1	0.68	0.11	0.865	0.0120
Bent Creek	0.883	0.0061	1.7	1.17	0.15	0.850	0.0235
Bethany	0.897	0.0034	0.2	0.88	0.81	0.893	0.0151
Bryson City	0.887	0.0041	-2.3	0.95	0.02	0.934	0.0198
Bushy Fork	0.901	0.0035	1.6	0.82	0.06	0.874	0.0147
Butner	0.911	0.0035	-0.3	0.83	0.69	0.917	0.0145
Castle Hayne	0.902	0.0040	-0.7	0.84	0.38	0.917	0.0169
Cherry Grove	0.899	0.0035	0.2	0.84	0.83	0.895	0.0149
Clemmons Middl	0.890	0.0036	-1.5	0.93	0.10	0.915	0.0160
County Line	0.891	0.0032	1.1	0.73	0.13	0.873	0.0120
Crouse	0.886	0.0033	-0.5	0.83	0.53	0.895	0.0141
Durham Armory	0.900	0.0036	-2.4	0.85	0.01	0.941	0.0152
Enoch	0.886	0.0032	0.4	0.81	0.58	0.879	0.0131
Franklinton	0.902	0.0037	-0.4	0.81	0.62	0.909	0.0148
Frying Pan	0.934	0.0037	2.7	1.11	0.02	0.887	0.0199
Fuquay	0.905	0.0034	-2.1	0.76	0.01	0.940	0.0133
Garinger HS	0.884	0.0032	2.5	0.71	0.00	0.846	0.0114
Golfview	0.908	0.0035	-2.5	0.79	0.00	0.950	0.0140
Hattie Ave. LP	0.893	0.0034	-0.7	0.80	0.40	0.904	0.0140
Jamesville	0.908	0.0039	-1.4	0.85	0.11	0.933	0.0165
Joanna Bald	0.933	0.0034	1.8	1.04	0.08	0.902	0.0177
Lenoir City	0.901	0.0037	-1.8	0.96	0.06	0.935	0.0180
Lenoir Communi	0.908	0.0038	-2.5	0.82	0.00	0.953	0.0154
Linville Falls	0.899	0.0040	-2.5	0.99	0.01	0.950	0.0202
Mendenhall	0.894	0.0033	0.0	0.77	0.97	0.893	0.0130
Millbrook	0.901	0.0036	-0.5	0.72	0.48	0.910	0.0131

Site	Ratio	SE	Intercept	SE	Prob (inter)	Linear Slope	SE
Mitchell	0.934	0.0033	5.0	1.30	0.00	0.852	0.0216
Mocksville	0.890	0.0035	1.0	0.80	0.21	0.872	0.0142
Monroe	0.900	0.0035	0.1	0.79	0.90	0.898	0.0140
Pitt County Ag	0.909	0.0035	-1.7	0.79	0.03	0.938	0.0139
Pittsboro	0.886	0.0040	-3.2	0.85	0.00	0.945	0.0165
Purchase Knob	0.926	0.0036	1.7	0.98	0.07	0.895	0.0179
Rockwell CSS	0.894	0.0033	0.5	0.77	0.49	0.885	0.0128
Shiloh Church	0.889	0.0036	0.8	0.89	0.35	0.875	0.0159
Tarboro	0.898	0.0037	-1.0	0.77	0.22	0.915	0.0142
Union Cross	0.887	0.0034	0.9	0.78	0.24	0.872	0.0134
Wade	0.898	0.0035	-1.4	0.75	0.06	0.923	0.0134
Waggin Trail R	0.908	0.0037	-3.1	0.99	0.00	0.965	0.0186
Waynesville El	0.895	0.0037	-0.8	0.92	0.41	0.909	0.0173
West Johnston	0.903	0.0035	-2.0	0.80	0.01	0.938	0.0143

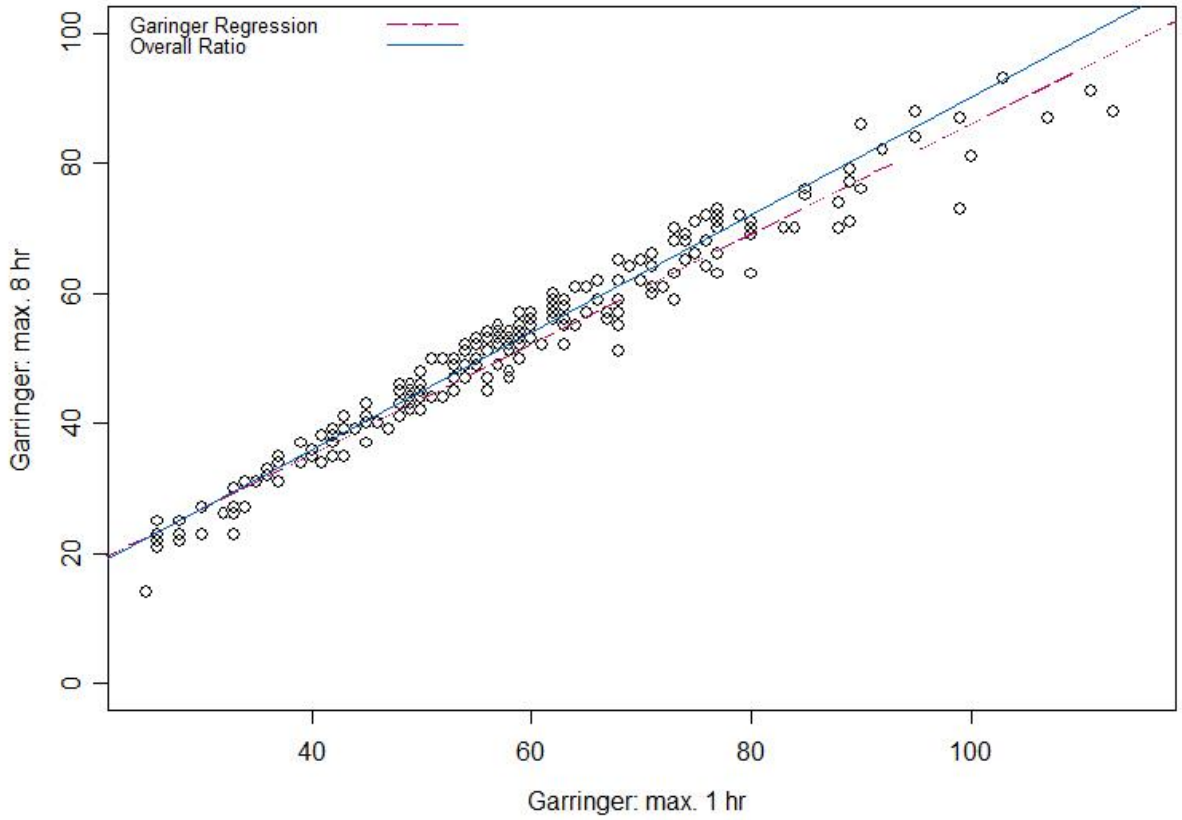

Figure 3. Surrogate Ozone Estimation for Garinger.
Scatter plot with Model 1 ratio and Model 4 regression lines overlaid. The Model 1 Overall Ratio line fits the data slightly better when 40 ppb < Max. 1 hour average < 80 ppb, while the Model 4 Regression line fits better when Max. 1 hour average > 80 ppb.