Atmospheric and Environmental Research

131 Hartwell Avenue
Lexington, Massachusetts
02421-3126
USA
Tel: +1 781 761-2288
Fax: +1 781 761-2299
www.aer.com

TECHNICAL MEMO

Investigating the Impact of Meteorology on O_3 and $PM_{2.5}$ Trends, Background Levels, and NAAQS Exceedances

Task 2: Effects of Meteorology on O_3 and $PM_{2.5}$ Trends

TCEQ Contract No. 582-15-50415
Work Order No. 582-15-54118-01
Deliverable 2.2
Revision 1.0

Prepared by:

Atmospheric and Environmental Research, Inc. (AER)
131 Hartwell Ave.
Lexington, MA 02466
Correspondence to: malvarad@aer.com

Prepared for:

Erik Gribbin
Texas Commission on Environmental Quality
Air Quality Division
Building E, Room 342S
Austin, Texas 78711-3087

June 30, 2015

Document Change Record

Revision	Revision Date	Remarks
1.0	**30 June 2015**	**Version submitted to TCEQ with deliverable**

Table of Contents

List of Figures

List of Tables

1 Introduction

This technical memo documents the files provided to TCEQ to complete Deliverable 2.2 of Work Order No. 582-15-54118-01. As stated in the Work Plan, this deliverable is:

Deliverable 2.1: Deliver a technical memo describing GLMs relating meteorological variables to measured MDA8 O_3 and $PM_{2.5}$ for urban areas in Table 1 based on data for the O_3 season (May through October) from 2005-2014 and $PM_{2.5}$ from 2005-2014. AER will also attach R scripts and other computer codes used to generate and/or analyze the GLMs.

Deliverable Due Date: June 30, 2015

Table 1. Urban areas of interest to this study.

Group 1 Urban Areas	Group 2 Urban Areas
Dallas/Fort Worth (DFW)	Beaumont/Port Arthur (BPA)
Houston/Galveston/Brazoria (HGB)	Tyler-Longview-Marshall (TLM)
San Antonio (SA)	
Austin/Round Rock (ARR)	

The GAMs and all associated data and scripts are in the gzipped tar file for the deliverable, which can be downloaded from the AER ftp server at:

`ftp://ftp.aer.com/pub/malvarad/p1952_deliverable_2_2_R1_0.tar.gz`

Our major findings are (see Sections 2.4, 2.5, and 2.6 for more details):

- The GAMs relating meteorological variables to the maximum MDA8 O_3 for each urban area generally explain 65-80% of the deviance (i.e. variability), consistent with the results of Camalier et al. (2007). The O_3 GAMs also generally show good fits with normally-distributed residuals and little dependence of the residual variance on the predicted value.
- In contrast, the GAMs relating meteorological variables to the maximum daily average $PM_{2.5}$ for each urban area only explain 30-40% of the deviance, and generally show much poorer fits with long, positive residual tails and a strong dependence of the variance of the residuals on the predicted value.
- Using meteorological predictors different from those listed in Camalier et al. (2007) can result in an improved GAM for MDA8 O_3 and daily average $PM_{2.5}$, but the improvement is less significant for $PM_{2.5}$.
- Two-fold cross validation analysis shows that the GAM fitting procedure results in GAMs that only perform slightly worse for the "test" data set as they do for the "training" data set, and thus the GAMs show little evidence of overfitting.
- However, the cross validation analysis also shows that the smooth function fit for some meteorological predictors can vary substantially depending on which half of the data is used to train the GAM. Thus the individual smooth functions from each GAM should be used with caution.

Section 2 of this memo briefly outlines the technical approach used to prepare the generalized additive models (GAMs) in the deliverable and Section 3 describes the files in the deliverable. Section 4 briefly outlines the quality assurance steps that have been performed. Further details and analysis of the results will be included in the project Final Report.

2 Technical Approach

As described in the Work Plan, AER derived updated GAMs for O_3 and $PM_{2.5}$ for selected monitoring sites within the urban areas in Table 1. For O_3, only data during the O_3 season (May to October) was analyzed, but $PM_{2.5}$ data for the entire year was analyzed.

AER first fit the data to the 8 meteorological parameters that were determined to give the best fit for urban O_3 by Camalier et al. (2007). As in that paper, a daily transport distance and transport direction were determined by 24-hour back-trajectories calculated with the HYSPLIT model (Draxler and Hess, 1997, 1998) driven with meteorology from the 32 km horizontal resolution North American Regional Reanalysis (NARR).

In addition to these "baseline" GAMs (referred to as "gam01_baseline" below and in the deliverable files), AER explored whether the addition or substitution of other meteorological variables significantly increased the amount of variability explained by the model. This resulted in two additional GAMs ("gam02_extended" and "gam03_extended") that are also included in the deliverable.

One of the dangers of using GAMs to perform the meteorological adjustment of pollutant trends is the possibility of "over-fitting," where some of the variability that is actually due to changes in air quality policy is accounted for in the GAM by the meteorological variables. AER explored the potential errors from over-fitting via cross validation. In cross validation, some of the data (the testing set) is removed before building the GLM. The remaining data (the training set) is used to derive the GAM parameters. The testing set can then be used to test the performance of the GAM in predicting "unseen" data (e.g., Starkweather et al., 2011).

Section 2.1 below describes the input data used to generate the GAMs, including a discussion of the processing we performed on the raw data to make it suitable for generating the GAMs. Section 2.2 describes the generation and evaluation of the HYSPLIT back trajectories. Section 2.3 gives an overview of our GAM fitting procedure, followed by an overview of the GAM results for both the baseline (Section 2.4) and extended (Section 2.5) GAMs. Section 2.6 then presents the results of the cross-validation analysis of the "gam03_extended" GAMs from Section 2.5.

2.1 Input Data and Processing

2.1.1 TCEQ Monitor Data

The TCEQ provided AER with air quality and meteorological monitoring data from the air quality monitoring network operated by the TCEQ, its grantees, or local agencies whose data is stored in the Texas Air Monitoring Information System (TAMIS) in and near the urban areas listed in Table 1 covering a ten-year period (2005-2014). AER then built Python scripts that processed the TCEQ air quality and meteorological data and calculate the average (daily, morning, afternoon, etc.) and derived quantities (e.g., deviations from 10-year monthly averages) needed for the GAM fitting. Following Camalier et al. (2007), these average and derived quantities for each urban area were calculated using a single surface site in the center of the urban area combined with the nearest radiosonde location available. The selected surface sites

for each urban region are given in Table 2 - they were selected to maximize the amount of data available at each site.

Table 2. Surface meteorological sites selected for GAM fitting.

Urban Area	Site #	Latitude (°)	Longitude (°)
Houston/Galveston/Brazoria	482011035	29.7337263	-95.2575931
Dallas/Fort Worth	484391002	32.8058183	-97.3565675
San Antonio	480290055	29.4072945	-98.431251
Austin/Round Rock	484530014	30.3544356	-97.7602554
Beaumont/Port Arthur	482450009	30.0364221	-94.0710606
Tyler-Longview-Marshall	481830001	32.3786823	-94.7118107

As noted in our Technical Memo for Deliverable 3.1 (R1.1, dated June 15, 2015), we developed a python script (*calc_bkgrd_ozone.py*, see Section 3.2) that calculated the MDA8 O_3 (ppbv) for all of the monitoring sites in the six urban areas. The MDA8 for a site was calculated as follows:

1. A running 8-hour average was calculated for each hour, averaged over that hour and the following seven hours. At least 6 hours in this 8-hour range had to have valid O_3 measurements for the 8-hour average to be considered valid.
2. The largest of each of the calculated 8-hour averages in a day was selected as the MDA8 for that day.
3. The maximum and minimum of the valid MDA8 O_3 values for all sites in the urban area were determined.
4. The minimum of the valid MDA8 O_3 values for the selected background sites were determined as the daily background concentration for that area.

A similar script (*calc_pm25.py*) was used to calculate daily average $PM_{2.5}$ values from the available hourly data. This average was calculated as follows:

1. If more than one $PM_{2.5}$ instrument was active for a site, the reported hourly values were averaged.
2. A daily average $PM_{2.5}$ value was then calculated for each site. At least 18 hours of that day had to have valid $PM_{2.5}$ measurements for the daily average to be considered valid.
3. The maximum and minimum of the valid $PM_{2.5}$ values for all sites in the urban area were determined.
4. The minimum of the valid $PM_{2.5}$ values for the selected background sites were determined as the daily background concentration for that area.

Two additional python scripts (*calc_GLM_all.py* and *calc_GLM_NCDC.py*) were used to calculate the potential meteorological predictors listed in Appendix A. The TCEQ monitor data, Integrated Global Radiosonde Archive data (IGRA, Section 2.1.2) and the integrated surface hourly (ISH) database of the National Climatic Data Center (NCDC, Section 2.1.3), along with the previously calculated MDA8 and $PM_{2.5}$ maximum and minimum concentrations and

parameter from the HYSPLIT back trajectories (Section 2.2), were merged by a final script (*merge_param_all_Camalier.py*). This script then outputs the final CSV file used in fitting the GAM model. These scripts are all described further in Section 3.

2.1.2 IGRA Radiosonde Data

The Integrated Global Radiosonde Archive (IGRA) provided upper atmosphere data used to derive the meteorological predictors for the GAMs. These data can be downloaded at ftp://ftp.ncdc.noaa.gov/pub/data/igra. Table 3 describes the sites selected for each urban area, which were selected because they were the closest sites to the center of each urban area that had continuous data for the 2005-2014 period. Section 3.1.1 describes these files in further detail.

Table 3. IGRA sites used for each urban area.

Urban Area	ID	Station Name	Lat. (°)	Lon. (°)
Houston/Galveston/Brazoria	72249	FORT WORTH	32.8	-97.3
Dallas/Fort Worth	72240	LAKE CHARLES	30.12	-93.22
San Antonio	72261	DEL RIO	29.37	-100.92
Austin/Round Rock	72261	DEL RIO	29.37	-100.92
Beaumont/Port Arthur	72240	LAKE CHARLES	30.12	-93.22
Tyler-Longview-Marshall	72248	SHREVEPORT	32.45	-93.83

2.1.3 NCDC Integrated Surface Hourly Data

We have also added data from the integrated surface hourly (ISH) database of the National Climatic Data Center (NCDC) to our dataset. We used the NCDC data to get estimates of surface pressure and relative humidity, as this data was not generally available in the TCEQ dataset. The NCDC sites used for each urban area are described in Table 4 below. These sites were selected because they were the closest sites to the center of each urban area that had continuous data for the 2005-2014 period. The dataset is described further in Section 3.1.2.

Table 4. NCDC surface sites used for each urban area.

Urban Area	USAF-WBAN_ID	Station Name	Lat. (°)	Lon. (°)
DFW	722590 03927	DALLAS/FT WORTH INTERNATIONAL	32.898	-97.019
HGB	722430 12960	G BUSH INTERCONTINENTAL AP/HOU	29.98	-95.36
SA	722530 12921	SAN ANTONIO INTERNATIONAL AIRP	29.544	-98.484
ARR	722544 13958	AUSTIN-CAMP MABRY ARMY NATIONA	30.321	-97.76
BPA	722410 12917	SOUTHEAST TEXAS REGIONAL AIRPO	29.951	-94.021
TLM	722470 03901	EAST TEXAS REGIONAL ARPT	32.385	-94.712

2.1.4 NARR Data

The North American Regional Reanalysis (NARR) meteorological data are available from 1979 to 2014 on a 3 hourly, 32 km grid. The NARR is an extension of the NCEP Global

Reanalysis but only for North America. Combining the higher resolution NCEP Eta Model (32km/45 layer) with a data assimilation system optimized for regional reanalysis results in better accuracy of the meteorological variables compared to the NCEP Global Reanalysis. The NARR data can be downloaded from the NOAA Air Resources Library (ARL) ftp server at ftp://arlftp.arlhq.noaa.gov/narr.

2.2 HYSPLIT Back Trajectories

We ran 24-hour HYSPLIT back-trajectories for each urban region for the 2005-2014 period. These back-trajectories were calculated using the 32 km horizontal resolution NARR, as these data were available in a form suitable to drive HYSPLIT for our entire study period (2005-2014), as opposed to the 12 km North American Mesoscale (NAM-12) data called for in the Work Plan, which were only available for 2008-2014. As in Camalier et al. (2007), these back-trajectories are calculated assuming an initial height of 300 m above ground level (AGL) and are started at noon local solar time. The starting points for the back-trajectories are the selected surface meteorological sites given in Table 2 above. The HYSPLIT model (Draxler and Hess, 1997, 1998) is available for download from the HYSPLIT website (http://ready.arl.noaa.gov/HYSPLIT.php). The performance of HYSPLIT driven with NARR meteorological fields was evaluated with tracer release studies by Hegarty et al. (2013).

The endpoints of the back-trajectories were used to calculate the 24-hour transport direction and distance for each urban area for the 2005-2014 period. This was done using the R functions *bearing* and *distMeeus* from the *geosphere* package (see the script *./hysplit_trajec/ calc_trajec.src*, described in Section 3.3.3). The function *bearing* gets the initial bearing (direction; azimuth) to go from point 1 to point 2 following the shortest path (a Great Circle). The function *distMeeus* calculates the shortest distance between two points (i.e., the 'great-circle-distance' or 'as the crow flies') using the WGS84 ellipsoid.

The HYSPLIT back-trajectories used in the model development appear reasonable and are generally consistent with the surface wind speed and direction measured near the center of each urban area. The HYSPLIT back-trajectory distance is generally correlated with the urban area average surface wind speed with a linear correlation coefficient (R) of 0.4-0.6. The frequency of both the daily average wind direction and the HYSPLIT back-trajectory bearings peak around 150^{o} (southeast, from the Gulf of Mexico) for all urban areas. However, the HYSPLIT back-trajectory bearings also show a secondary maximum at 0^{o} (north) not seen in the daily average wind directions.

We also examined a few ensemble back-trajectories, initialized from slightly different locations, to determine the potential uncertainty of the back-trajectory calculations. Figure 1 shows an example ensemble back-trajectory calculation for August 25, 2013 in Houston/Galveston/Brazoria, a day of high MDA8 O_3. We can see that the back-trajectories all follow a consistent qualitative shape, although the exact locations of the end points can differ. These results give us confidence that our HYSPLIT results are representative of the air masses entering the urban areas, but that differences in distance of less than ~100 km and differences in bearing of less than ~20^{o} are unlikely to be significant.

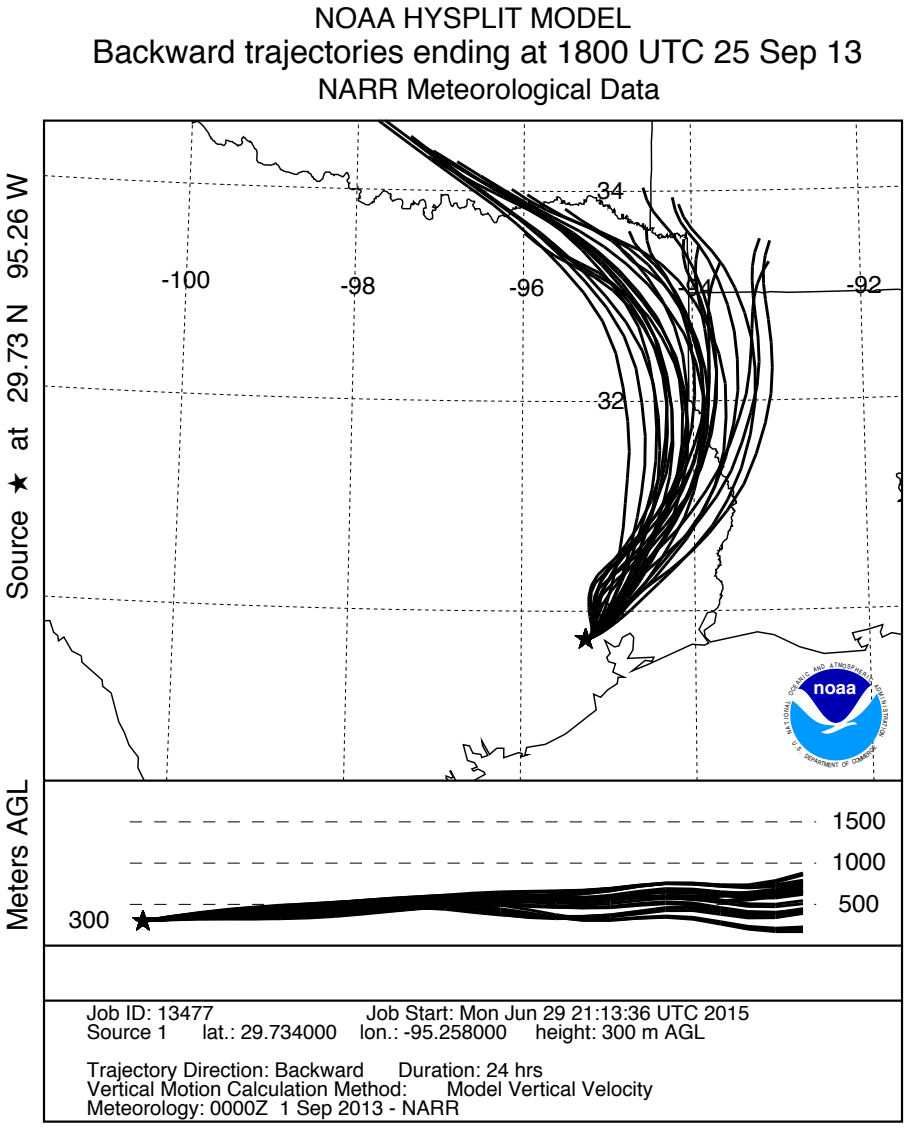

Figure 1. Ensemble back-trajectory run for the Houston/Galveston/Brazoria area on August 25, 2013.

2.3 Generalized Additive Model (GAM) Fitting Procedure

In our procedure, we fit the maximum MDA8 O_3 value and the maximum 24-hour average $PM_{2.5}$ value for each urban area using the GAM modeling function in the *mgcv* package in R (Wood, 2006). The GAM can be written as follows:

$$g(\mu_i) = \beta_o + f_1(x_{i,1}) + f_2(x_{i,2}) + \cdots f_n(x_{i,n}) + f_p(D_i) + W_d + Y_k$$

where i is the ith day's observation, $g(\mu_i)$ is the "link" function (here, a log link is used), $x_{i,j}$ are the n meteorological predictors fit, with the corresponding $f_j(x_{i,j})$ being a (initially unknown) smooth function of $x_{i,j}$ made from a cubic-spline basis set. Following Camalier et al. (2007),

three non-meteorological predictors are also included: a smooth function $f_p(D_i)$ of the Julian day of the year (D_i); a factor for the day of the week W_d and a factor for the year Y_k. As we are only fitting O_3 data during the O_3 season (May-October), $f_p(D_i)$ is built with a non-periodic cubic spline basis for O_3, but for $PM_{2.5}$, a periodic cubic spline basis is used. To reduce the possibility of over-fitting the data, we set the "gamma" parameter to 1.4 for these fits, as recommended by Wood (2006).

We also added an automated process to determine if a predictor that is not significant at the α = 0.001 level could be eliminated from the fit without significantly degrading the performance of the model. In this process, the meteorological predictor with the highest *p* value is removed and a second GAM is fit. This is then compared to the original model using the ANOVA procedure in R. If the second model with the variable removed is not different from the original model at the α = 0.01 level, the variable is "dropped" from the fit and the variable with the next highest p value is tested. If the second model is significantly worse than the original model, the variable is kept and no other variables are tested or dropped. Because of this, although the GAMs for a given pollutant may start with the same predictors for all urban areas, the final GAM selected may have different predictors depending on which variables were dropped for each urban area.

2.4 Baseline GAMs (gam01_baseline)

2.4.1 Description

We have developed "baseline" GAMs for the maximum MDA8 O_3 and daily average $PM_{2.5}$ in each area, where we use the eight meteorological parameters identified as significant by Camalier et al. (2007) in their study of O_3 in eastern US cities. These parameters are listed in Table 5 below. The automated process to remove insignificant predictors was not used for these fits.

Table 5. Meteorological parameters used in the "baseline" GAMs. The column name is given in italics.

Daily maximum temperature (^{o}C, *daily_max_T*)
Mid-day average (10 am–4 pm average) relative humidity (%, *NCDC.Mid.day.RH*)
Morning (7–10 am) average wind speed (m s^{-1}, *morning_ws*)
Afternoon (1–4 pm) average wind speed (m s^{-1}, *afternoon_ws*)
Morning surface temperature difference (1200 UTC) (temperature at 925 mb–temperature at surface at 1200 UTC) (^{o}C, *T_diff_925mb*)
Deviation in 1200 UTC temperature of 850 mb surface from 10-year monthly average (^{o}C, *T_dev_850mb*)
Transport direction (degrees clockwise from North, *HYSPLIT_DIST..m.*)
Transport distance (m, *HYSPLIT_DIST..m.*)

2.4.2 Results

To illustrate the results, we discuss the baseline GAM fits for Houston in detail. Similar plots for all urban areas are contained in the deliverable as described in Section 3.6. Figure 2 shows the smooth functions from the baseline GAM fit of the natural logarithm of the HGB maximum

MDA8 O_3 values to the meteorological predictors in Table 5. 95% confidence intervals are shown in red. The periodic day of year function is also shown. This model explains 74% of the deviance of the MDA8 O_3 values. This is consistent with the Camalier et al. (2007) results, which showed the predictive power of their models (measured by the R^2 statistic) to be between 0.56 and 0.80 for the cities in that study. In this case, all eight meteorological predictors and the day-of-year function are statistically significant at the $\alpha = 0.001$ level. As expected, the model fit shows O_3 generally increasing with daily maximum temperature, decreasing with RH, decreasing with wind speed, and increasing with vertical stability (positive values of T_diff_925mb). In addition, the predicted O_3 mixing ratio drops when the wind is from the southeast, as expected for air blowing from the Gulf of Mexico to Houston. The day-of-year function may reflect the fact that the mean mixing height increases in the summer, leading to a decrease in MDA8 O_3 in the middle of the ozone season. For the weekday factor variables, the largest average MDA8 values are Wednesday-Friday, with Sunday having the lowest average MDA8 values, as expected. The differences between Sunday and the Wednesday-Friday period are significant at the $\alpha = 0.001$ level.

The year-to-year differences in the meteorologically-adjusted natural logarithm of MDA8 O_3 are shown in Figure 3. All of the differences from the base year of 2005 are statistically significant at the $\alpha = 0.001$ level except for 2006. However, the two-fold cross-validation tests (described in Section 2.6 below) show that the year-to-year changes in MDA8 O_3 determined with different randomly-distributed halves of the dataset can give very different results (the red and blue circles in Figure 3), although these are generally within the 95% confidence interval of the original fit. It is also unclear why there would be a sudden increase between 2010 and 2011 that is not accounted for by the meteorological predictors. Thus, while we can be reasonably confident that the meteorologically-adjusted MDA8 O_3 for Houston in 2014 was significantly lower than in 2005, the magnitude and shape of the trend over the years is less certain.

The standard GAM evaluation plots (made with the *gam.check* function in R) for this case are shown in Figure 4. These plots indicate a good fit, as the model residuals are roughly normally distributed and show no trend versus predicted value. The variance of the residuals is lower for low values of the predictor, but this reflects the fact that the measured MDA8 O_3 values cannot go below 0.

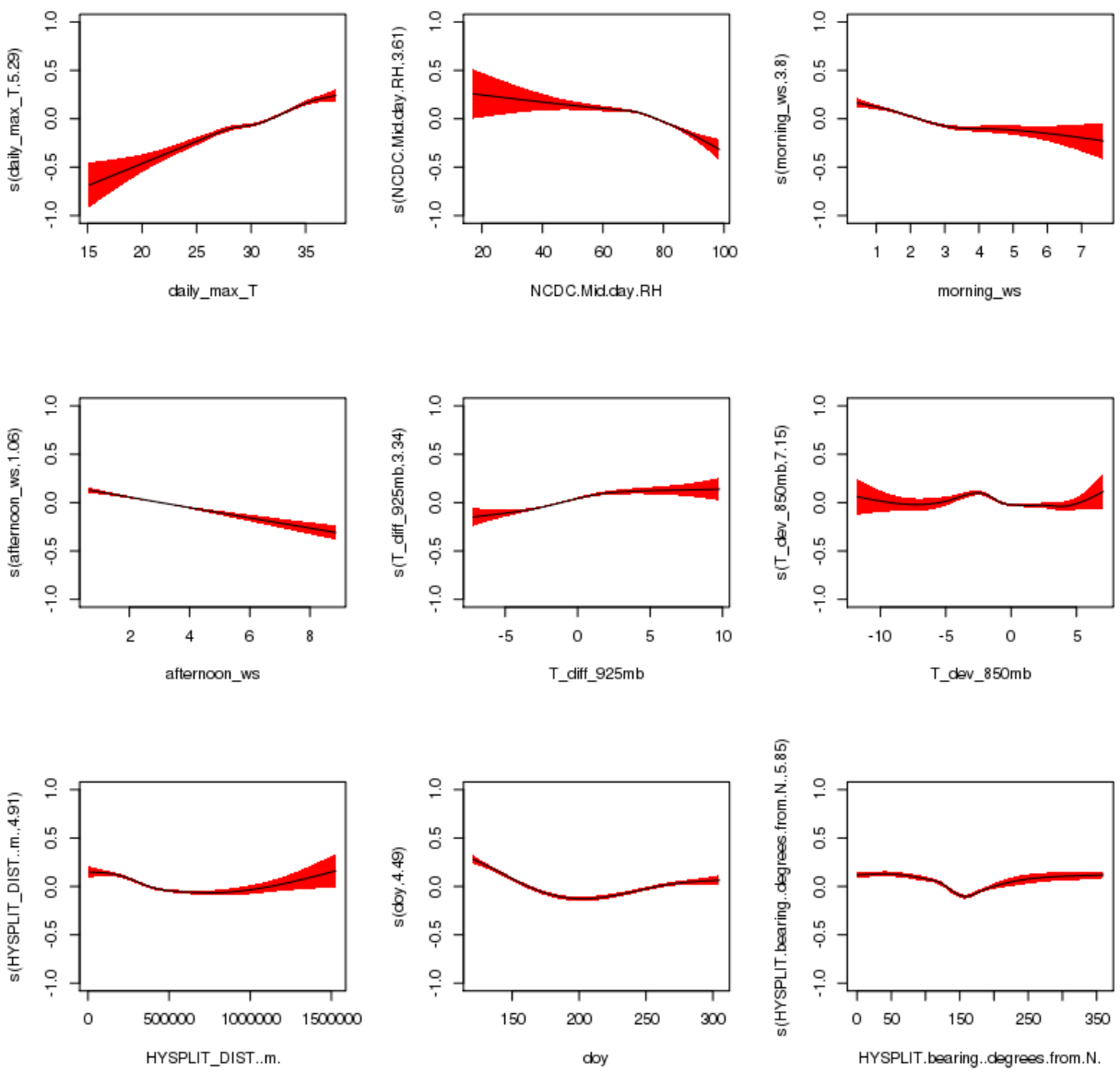

Figure 2. Smooth functions for the baseline GAM (gam01_baseline) fit to HGB MDA8 O_3 data. The y-axis scale is the scale of the "linear predictor", i.e. the deviation of the natural logarithm of the MDA8 O_3 in ppbv from its mean value.

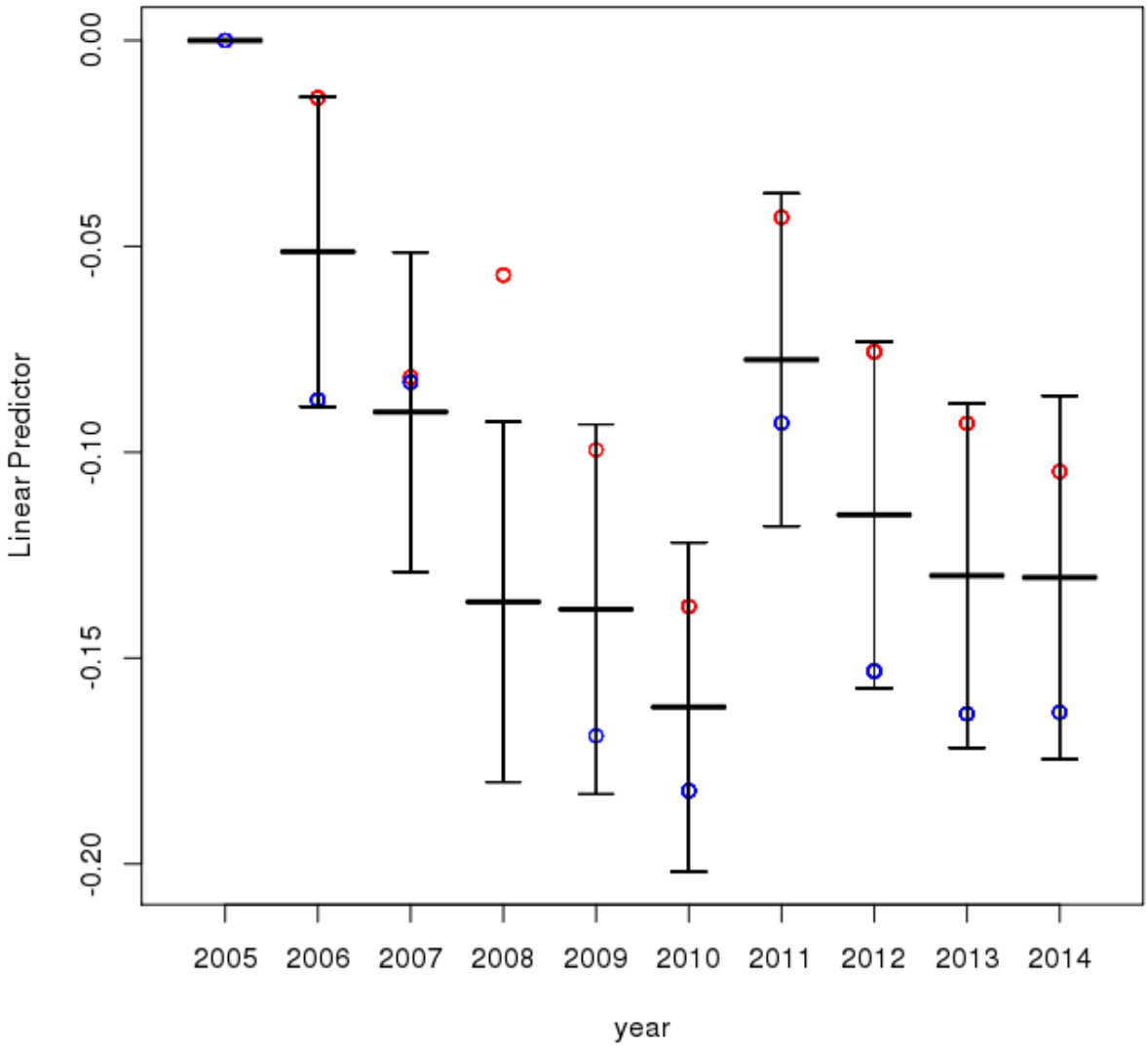

Figure 3. Year-to-year deviations from 2005 for the baseline GAM (gam01_baseline) fit to HGB MDA8 O_3 data. The y-axis scale is the scale of the "linear predictor", i.e. the deviation of the natural logarithm of the MDA8 O_3 in ppbv from its mean value. The black center bar is the mean value while the error bars are the 95% confidence intervals. The red and blue circles are the mean values from the two-fold cross-validation analysis of Section 2.6.

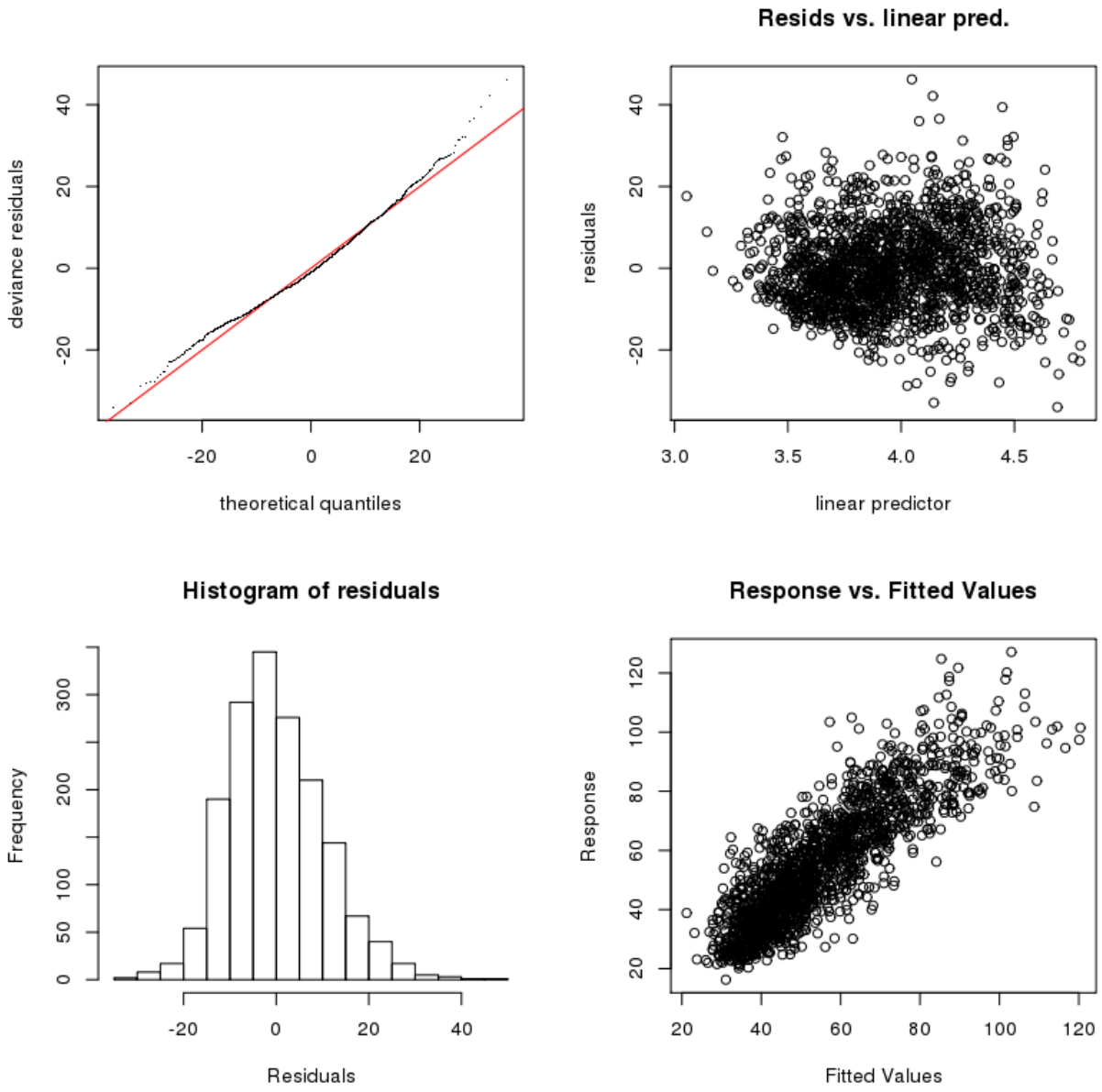

Figure 4. GAM evaluation plots for the baseline GAM (gam01_baseline) fit to HGB MDA8 O_3 data.

Figure 5 shows the smooth functions from the baseline GAM fit of the natural logarithm of the HGB maximum daily average $PM_{2.5}$. This model only explains 38% of the deviance in the $PM_{2.5}$ values, and so the baseline meteorological parameters in Table 5 give a much poorer prediction than the same parameters do for O_3. Again, all eight meteorological predictors and the day-of-year function are statistically significant at the $\alpha = 0.001$ level. Like O_3, increasing maximum temperature generally leads to increasing $PM_{2.5}$. However, in this case there is an indication that at the highest temperatures this relationship may not hold, possibly because evaporation of semi-volatile organic and ammonium nitrate aerosol begins to compete with the

increased chemical production of secondary aerosol with increasing temperature. The impacts of wind speed are much less strong as well, potentially reflecting increased dust and marine aerosol emission at high wind speeds. Similarly, the impact of air blowing from the Gulf is less pronounced for $PM_{2.5}$, perhaps reflecting the increased transport of marine aerosol to Houston during these periods. All weekdays have larger $PM_{2.5}$ values than Sunday, and with the exception of Saturday these differences are significant at the $\alpha = 0.001$ level.

The year-to-year differences in the meteorologically-adjusted natural logarithm of daily average $PM_{2.5}$ are shown in Figure 6. The years 2009 to 2014 are all significantly lower than 2005 at the $\alpha = 0.001$ level. The two-fold cross-validation tests (described in Section 2.6 below) show little difference in the observed trend, with both randomly-distributed halves of the dataset showing slight increases in 2006 and 2007 followed by dramatic decreases.

The GAM evaluation plots in Figure 7 indicate a poorer fit for $PM_{2.5}$ than for O_3, as the residuals show a long positive tail and the variance of the residuals is a strong function of the value of the linear predictor.

Table 6 below summarizes the percentage of the deviance explained by the baseline GAMs for each urban area for MDA8 O_3 and daily-average $PM_{2.5}$. For O_3, the values vary between 65.7% (San Antonio) and 73.9% (Houston/Galveston/Brazoria), similar to the range of 56-80% reported by Camalier et al. (2007). The performance for $PM_{2.5}$ is much poorer for all urban areas, with values between 30.0% (Beaumont/Port Arthur) and 37.8% (Houston/Galveston/Brazoria).

The generalized cross validation (GCV; see p.132 of Wood, 2006) score and Akaike's Information Criterion (AIC; see p.68 of Wood, 2006) for each GAM is also shown in Table 6. Both of these criteria attempt to compensate for the fact that adding redundant parameters to a model will always increase the likelihood of the model (and the amount of deviance explained), even if the new parameters are only "modeling the noise" of the data, i.e., over-fitting the data. For a given urban area and pollutant, the model with the lower GCV score and AIC is considered to be a better fit for the data. These scores will be compared to the values from the extended GAMs discussed in Section 2.6.

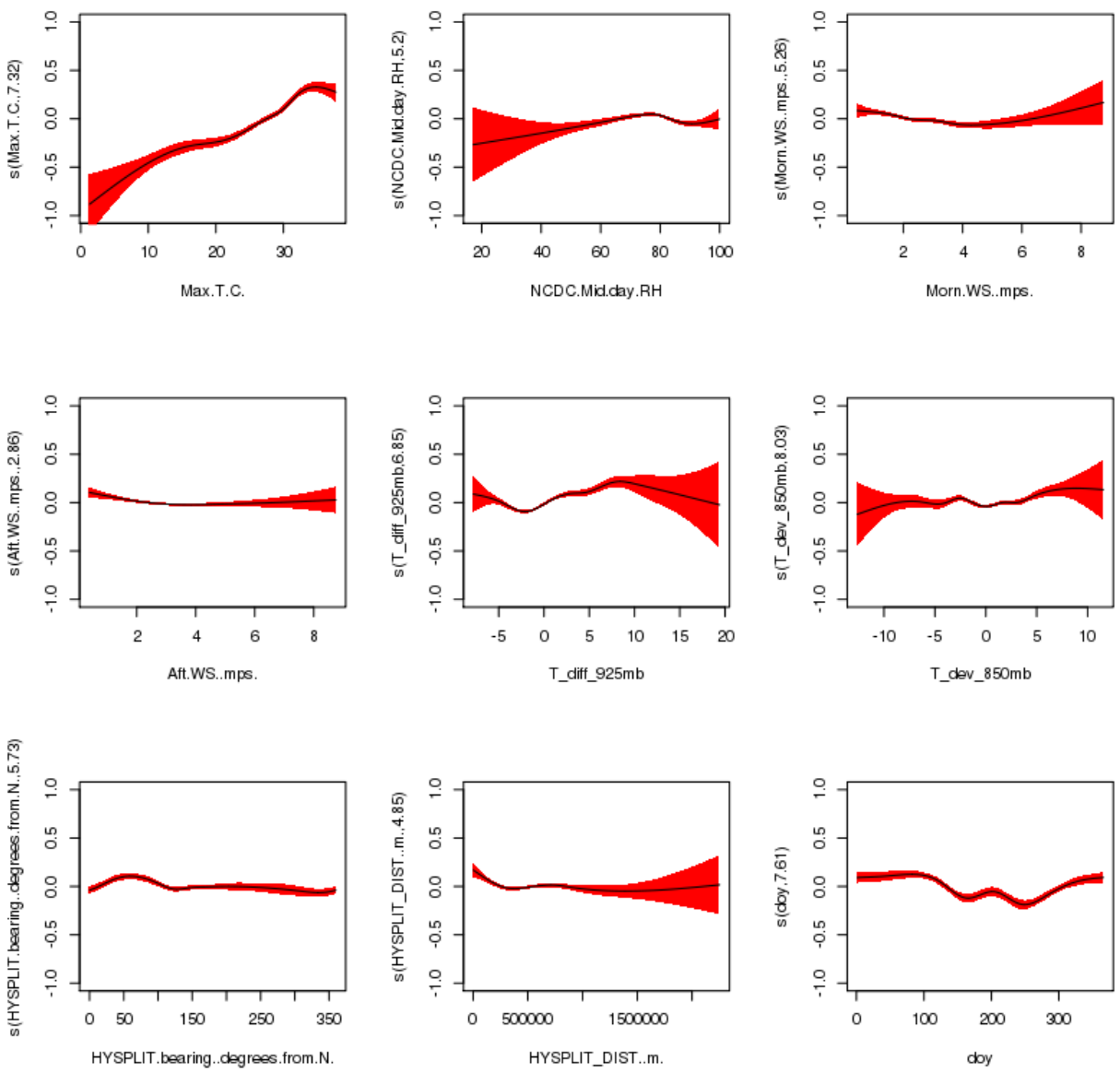

Figure 5. Smooth functions fit for the baseline GAM (gam01_baseline) fit to HGB daily average $PM_{2.5}$ data. The y-axis scale is the scale of the "linear predictor", i.e. the deviation of the natural logarithm of the daily average $PM_{2.5}$ in µg m^{-3} from its mean value.

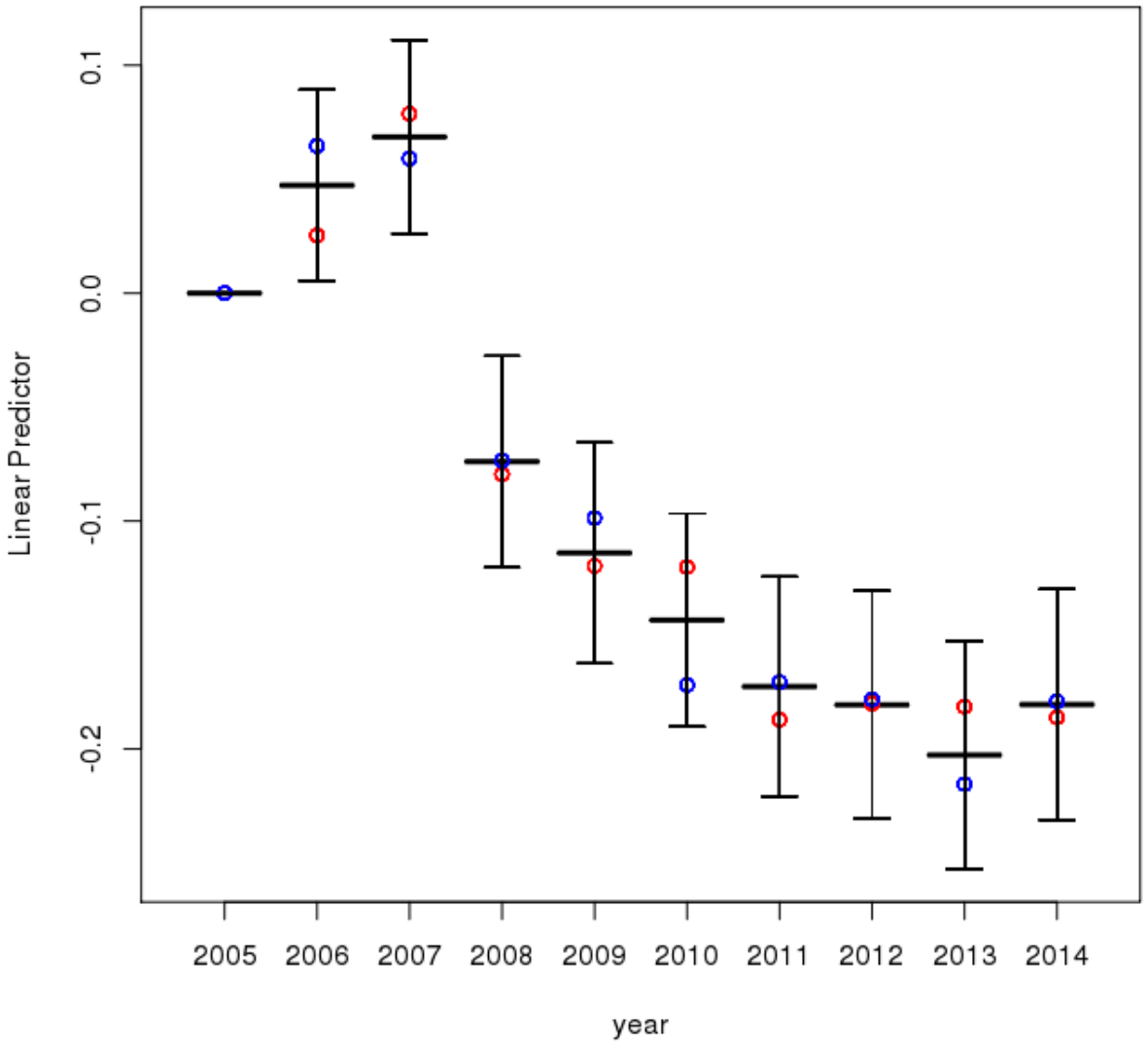

Figure 6. Year-to-year deviations from 2005 for the baseline GAM (gam01_baseline) fit to HGB daily average $PM_{2.5}$ data. The y-axis scale is the scale of the "linear predictor", i.e. the deviation of the daily average $PM_{2.5}$ in µg m^{-3} from its mean value. The black center bar is the mean value while the error bars are the 95% confidence intervals. The red and blue circles are the mean values from the two-fold cross-validation analysis of Section 2.6.

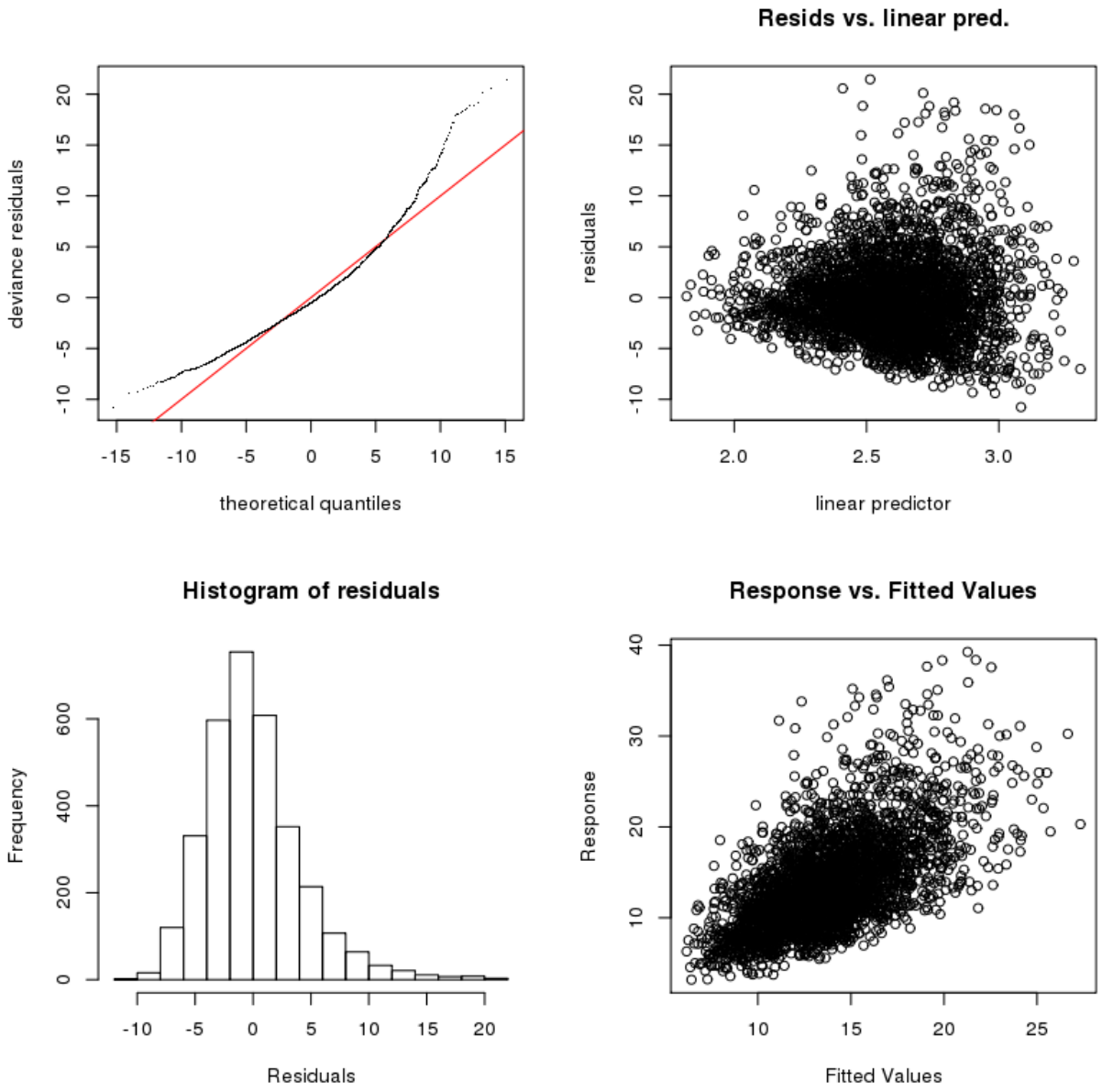

Figure 7. GAM evaluation plots for baseline GAM (gam01_baseline) fit to HGB MDA8 O_3 data.

Table 7 lists the meteorological predictors in each urban area that were not significant at the α=0.001 level for maximum MDA8 O_3 and maximum daily average $PM_{2.5}$. In addition, we examined the smooth functions fit for each predictor for similarities and differences between the urban areas. For maximum MDA8 O_3:

- The daily maximum temperature functions all show increasing O_3 with increasing temperature, but the fits for DFW, SA, and ARR become flat for temperatures > 30 ^oC, while the other areas show no such flattening off.

- The mid-day RH functions all show decreasing O_3 with increasing RH, and have a similar shape for all urban areas (relatively flat until 60% RH, then increasing at higher RH).
- O_3 decreases with morning wind speed for all urban areas except ARR (where it is fairly flat).
- O_3 either decreases with afternoon wind speed or the predictor is not significant.
- All urban areas except DFW show increasing O_3 with increasing stability (*T_diff_925mb*). The predictor is fairly flat for DFW with maxima at either end that may not be significantly different from zero.
- The deviation of the 850 mbar temperature from the monthly average (*T_dev_850mb*) is insignificant for ARR and SA, and may just be fitting noise for the other urban areas as there is little consistency in the functional forms.
- O_3 decreases with HYSPLIT back-trajectory distance up to ~1000 km, at which point it becomes highly uncertain due to the low number of points, but may begin to increase.
- All the urban areas show a drop in O_3 at a HYSPLIT bask-trajectory bearing of ~150^o (southeast), likely due to reduced background O_3 from flows from the Gulf of Mexico.
- The day-of-year function shows a minimum at ~200 Julian days (July) in each urban area.

For maximum daily average $PM_{2.5}$:

- All urban areas generally show $PM_{2.5}$ increasing with daily maximum temperature, but the effect is fairly weak for ARR, and SA, DFW, and HGB suggest that the trend flattens out or reverses at temperatures > ~30 ^{o}C.
- The fits for mid-day RH are very uncertain at low (< 40%) and high (> 80%) values, and the functional shape changes significantly between urban areas, with SA and ARR generally showing decreasing $PM_{2.5}$ with increasing RH, HGB and BPA showing an opposite trend, and TLM showing a maximum around 70% RH.
- $PM_{2.5}$ either trends down with increasing morning wind speed or the effect is insignificant.
- $PM_{2.5}$ generally trends down with increasing afternoon wind speed, but HGB, DFA, and BPA show a highly uncertain upward trend for wind speeds greater than 6 m/s.
- All urban areas show increasing $PM_{2.5}$ with increasing stability (*T_diff_925mb*), but the effect is fairly weak for TLM.
- $PM_{2.5}$ generally trends upward with increasing deviation of the 850 mbar temperature from the monthly average (*T_dev_850mb*).
- $PM_{2.5}$ decreases with HYSPLIT back-trajectory distance up to ~500 km, at which point it becomes flatter and highly uncertain due to the low number of points.
- All urban areas show a maximum for $PM_{2.5}$ around a HYSPLIT back-trajectory bearing of ~60^o (northeast) and a minimum around 320^o (northwest), possibly due to the relative difference in the $PM_{2.5}$ concentrations in the western and eastern US. Most urban areas also show a secondary minimum around ~150^o (southeast), likely due to flows from the Gulf of Mexico.
- The day-of-year functions for all urban areas are lower in the summer, likely reflecting the higher mixing heights in this season. The maximum is generally between 50-100

Julian days (around March), and ARR, SA, and HGB show a secondary maximum at ~200 Julian days (July).

Table 6. Deviance explained by the baseline GAMs (gam01_baseline) for each urban area and pollutant and the corresponding GCV and AIC values.

Urban Area	MDA8 O_3			Daily-Average $PM_{2.5}$		
	Deviance Explained (%)	GCV	AIC	Deviance Explained (%)	GCV	AIC
DFW	72.8	87.71	13,060	35.2	17.05	19,360
HGB	73.9	118.2	12,680	37.8	18.28	19,010
SA	65.7	80.80	13,080	33.6	16.07	19,040
ARR	66.6	73.63	12,730	34.0	15.19	19,200
BPA	71.2	93.68	12,640	30.0	23.65	20,160
TLM	70.4	70.70	12,670	35.8	16.87	19,210

Table 7. Meteorological predictors that were not significant at the α=0.001 level for the baseline GAMs (gam01_baseline).

Urban Area	MDA8 O_3	Daily-Average $PM_{2.5}$
DFW	*None*	*NCDC.Mid.day.RH, afternoon_ws, HYSPLIT_DIST..m.*
HGB	*None*	*None*
SA	*T_dev_850mb*	*morning_ws*
ARR	*T_dev_850mb*	*morning_ws*
BPA	*morning_ws*	*morning_ws*
TLM	*afternoon_ws*	*T_dev_850mb, afternoon_ws*

2.5 Extended GAMs (gam02_extended and gam03_extended)

2.5.1 Description

We explored whether a different set of meteorological predictors than those used by Camalier et al. (2007) and used in the baseline GAMs of Section 2.4 could provide a better fit to the maximum MDA8 O_3 and maximum daily average $PM_{2.5}$ for each urban area. We used a three-step procedure to select an appropriate subset of meteorological predictors for these extended GAMs.

First, a large set of potential meteorological predictors was assembled from the TCEQ, IGRA, and NCDC ISH data described in Section 2.1, as well as the HYSPLIT back-trajectory endpoints described in Section 2.2. The 60 potential predictors in Camalier et al. (2007) were

used to guide the assembly of this set. The final files containing these predictors are described in Section 3.4.2, and predictors in those files are listed in Appendix A and in the file *./csv_files/final_files/GAMparam_readme.txt* in the deliverable.

Second, the meteorological predictors were screened to remove combinations of variables that were both (a) highly correlated with each other and (b) likely represented the same physical quantity. Highly correlated variables generally represent the same information, and including both of them in the GAM can cause problems, just as including two nearly identical variables in a linear fit can result in arbitrarily large, unconstrained values of the slopes for each variable. In this step, we focused on identifying the true number of reasonably independent (uncorrelated) variables that best correlated with the maximum MDA8 O_3 and daily average $PM_{2.5}$ for each urban area. For example, of the four initial surface temperature variables (maximum, morning average, afternoon average, and diurnal change), it was found that the first three were highly correlated with each other ($R > 0.8$). This is to be expected, as the maximum temperature will generally happen in the afternoon, and days with hot afternoons generally have hot mornings as well. Thus we conclude that there are only two independent surface temperature variables in that set, one representing an effective maximum temperature and one representing the diurnal temperature change. As the mean afternoon temperature was most correlated with MDA8 O_3 and daily average $PM_{2.5,}$ it was selected to represent the effective maximum temperature in the extended GAM fits. Similar analyses were performed for the variable sets representing humidity, combinations of temperature and humidity (e.g., dew point temperature and apparent temperature), surface wind speed and direction, upper air temperature, and pressure/geopotential height.

Third, the variables that passed the correlation screening described above were used to form initial GAMs for each urban area and pollutant. This would occasionally reveal additional variables that appeared to be strongly linked, such that the smooth function fit to each variable would have a very large uncertainty, and the two members of the pair would have opposing (cancelling) effects. In these cases, one member of the pair was removed and the fit run again.

The selected meteorological predictors for maximum MDA8 O_3 are listed in Table 8 while the predictors for maximum daily average $PM_{2.5}$ are listed in Table 9. These predictors were used to fit the "large" extended GAMs (gam02_extended). These fits did use the automated selection procedure described in Section 2.3 to remove insignificant predictors. Analysis of the final GAMs showed that some predictors were either dropped or not significant at the $\alpha = 0.001$ level for 4 or more of the urban areas. Thus, these predictors were removed and an additional "small" extended GAM fit was performed (gam03_extended). The variables removed from these fits are indicated at the bottom of Tables 8 and 9. Note for Tyler-Longview-Marshall, the large and small extended GAM fits are identical, as the variables removed for the small extended GAM were also removed from the large extended GAM by the automated selection procedure.

Table 8. Meteorological parameters used in the extended MDA8 O_3 GAMs

Meteorological Variable	Column Name	In gam03?
Afternoon (1–4 pm) mean temperature (^oC)	*afternoon_mean_T*	Yes
Diurnal temperature change (^oC)	*diurnal_T*	Yes
Daily average relative humidity (%)	*NCDC.Avg.RH*	Yes
Daily average dew point (^oC)	*NCDC.Avg.Dew.Point..C.*	Yes
Daily average wind speed (m s^{-1})	*daily_ws*	Yes
Daily average wind direction (degrees clockwise from North)	*daily_wd*	Yes
Morning surface temperature difference (1200 UTC) (temperature at 850 mbar – temperature at surface at 1200 UTC) (^oC)	*T_diff_850mb*	Yes
Transport direction (degrees clockwise from North)	*HYSPLIT.bearing..degrees.from.N.*	Yes
Transport distance (km)	*HYSPLIT_DIST..m.*	Yes
Deviation in 1200 UTC temperature of 850 mbar surface from 10-year monthly average (^oC)	*T_dev_850mb*	**NO**
Geopotential Height at 850 mbar and 1200 UTC (m)	*GH_850.m.*	**NO**
Surface solar radiation (Langy/min)	*SolarRadiation.Langy.min.*	**NO**

Table 9. Meteorological parameters used in the extended daily average $PM_{2.5}$ GAMs

Meteorological Variable	Column Name	In gam03?
Afternoon (1–4 pm) mean temperature (^oC)	*afternoon_mean_T*	Yes
Daily average relative humidity (%)	*NCDC.Avg.RH*	Yes
Temprature at 925 mbar and 1200 UTC (^oC)	*T_925mb*	Yes
Daily average wind speed (m s^{-1})	*daily_ws*	Yes
Morning surface temperature difference (1200 UTC) (temperature at 850 mbar – temperature at surface at 1200 UTC) (^oC)	*T_diff_850mb*	Yes
Transport direction (degrees clockwise from North)	*HYSPLIT.bearing..degrees.from.N.*	Yes
Transport distance (km)	*HYSPLIT_DIST..m.*	Yes
Surface solar radiation (Langy/min)	*SolarRadiation.Langy.min.*	Yes
Deviation in 1200 UTC temperature of 850 mbar surface from 10-year monthly average (^oC)	*T_dev_850mb*	**NO**
Diurnal temperature change (^oC)	*diurnal_T*	**NO**
Daily average wind direction (degrees clockwise from North)	*daily_wd*	**NO**

2.5.2 Results

Table 10 summarizes the percentage of the deviance explained by the large extended GAMs for each urban area for MDA8 O_3 and daily-average $PM_{2.5}$, while Table 11 shows the same for the small extended GAMs. The tables show that the large extended GAMs (gam02_extended) give slightly better fits than the small extended GAMs (gam03_extended). However, this difference is fairly small, and an examination of the smooth fits for the variables contained in each GAM show little difference in the functional shape. Despite the lower GCV and AIC scores, it seems likely that the additional predictive power from the large extended GAMS over the small is mainly from having an additional three variables to use to fit the noise.

For maximum MDA8 O_3, both extended GAMs are clear improvements over the baseline GAMs described in Section 2.4, as indicated both by the larger percentage of deviance explained (range of 74-79% versus 65-74%) and the lower GCV and AIC scores. For maximum daily average $PM_{2.5}$, the improvement is less clear, with only two urban areas (Dallas/Fort Worth and Houston/Galveston/Brazoria) showing both lower GCV and AIC scores in the small extended GAMs than in the baseline GAM.

Based on these results, we recommend using the small extended GAMs (gam03_extended) for most purposes, with the baseline GAMS (gam01_baseline) mainly used for comparison with the results of Camalier et al. (2007). In the rest of Section 2, we focus on the small extended GAMs (gam03_extended). However, all three sets of GAMs are included in the deliverable for completeness.

Table 10. Deviance explained by the large extended GAMs (gam02_extended) for each urban area and pollutant and corresponding GCV and AIC values.

Urban Area	MDA8 O_3			Daily-Average $PM_{2.5}$		
	Deviance Explained (%)	GCV	AIC	Deviance Explained (%)	GCV	AIC
DFW	78.7	69.51	12,160	39.2	16.22	18,800
HGB	79.3	97.46	11,480	40.9	17.42	17,800
SA	76.1	57.31	12,390	36.1	15.55	19,030
ARR	75.0	55.81	12,210	36.4	14.69	19,320
BPA	76.1	79.19	12,380	30.9	23.49	20,120
TLM	73.7	63.21	12,600	37.5	16.37	19,390

Table 11. Deviance explained by small extended GAMs (gam03_extended) for each urban area and pollutant and corresponding GCV and AIC values.

Urban Area	MDA8 O_3			Daily-Average $PM_{2.5}$		
	Deviance Explained (%)	GCV	AIC	Deviance Explained (%)	GCV	AIC
DFW	78.2	70.47	12,350	38.2	16.27	19,100
HGB	78.6	98.52	12,520	38.8	18.04	18,120
SA	75.6	58.46	12,560	34.8	15.74	19,080
ARR	74.5	56.33	12,370	34.2	15.07	19,420
BPA	75.5	79.84	12,550	30.1	23.73	20,380
TLM	73.7	63.21	12,600	37.5	16.37	19,390

Similar to Section 2.4, we discuss the small extended GAMs for Houston in detail to illustrate the results. Similar plots for all urban areas are contained in the deliverable as described in Section 3.6. Figure 8 shows the smooth functions from the small extended GAM fit of the natural logarithm of the HGB maximum MDA8 O_3 values to the meteorological predictors. The periodic day of year function is also shown, and the 95% confidence intervals are shown in red. All meteorological predictors used in gam03_extended were significant at the $\alpha = 0.001$ level except for average relative humidity, but that predictor was not removed by the automated selection procedure. As expected, the model fit shows O_3 generally increasing with daily maximum temperature, decreasing with increased humidity (increasing RH and dew point temperature), decreasing with wind speed, and increasing with vertical stability (positive values of *T_diff_850mb*). In addition, the predicted O_3 mixing ratio drops when the wind is from the southeast, as expected for air blowing from the Gulf of Mexico to Houston. The day-of-year

function is generally decreasing through the ozone season. For the weekday factor variables, the largest average MDA8 values are Wednesday-Friday, similar to the baseline GAM results. The differences between Sunday and the Wednesday-Friday period are significant at the $\alpha = 0.001$ level.

The year-to-year differences in the meteorologically-adjusted natural logarithm of MDA8 O_3 are shown in Figure 9. All of the differences from the base year of 2005 are statistically significant at the $\alpha = 0.001$ level except for 2007. The sudden change between 2010 and 2011 seen in the baseline GAM (Figure 3) is now gone, so that there is now a sharp decrease from 2007 to 2008 followed by a gradual (but not statistically significant) increase.

The standard GAM evaluation plots for this case are shown in Figure 10. These plots indicate a good fit, as the model residuals are roughly normally distributed and show no trend versus predicted value.

Figure 11 shows the smooth functions from the baseline GAM fit of the natural logarithm of the HGB maximum daily average $PM_{2.5}$. All eight meteorological predictors and the day-of-year function are statistically significant at the $\alpha = 0.001$ level. Like O_3, increasing maximum temperature generally leads to increasing $PM_{2.5}$. However, as in the baseline GAM there is an indication that at the highest temperatures this relationship may not hold, possibly because of the evaporation of semi-volatile aerosol components. Increasing wind speed tends to decrease $PM_{2.5}$ at low values (0-4 m/s) but appears to increase $PM_{2.5}$ at higher values (4-7 m/s), possibly reflecting increased dust and marine aerosol emission with higher wind speeds. Similarly, the impact of air blowing from the Gulf is less pronounced for $PM_{2.5}$, perhaps reflecting the increased transport of marine aerosol to Houston during these periods. $PM_{2.5}$ increases with increased vertical stability (positive values of *T_diff_850mb*) as expected. The negative dependence on solar radiation may reflect that lower values of solar radiation are seen on cloudy days, and SO_2 is rapidly oxidized to aerosol sulfate within clouds. The day-of-year dependence is consistent with an increase in the mean mixing layer height in the summer, leading to relatively lower values of $PM_{2.5}$ on those days. For the weekday factor variables, the largest daily average $PM_{2.5}$ values are Tuesday-Friday, with Sunday having the lowest values, as expected. The differences between Sunday and the Tuesday-Friday period are significant at the $\alpha = 0.001$ level.

The year-to-year differences in the meteorologically-adjusted natural logarithm of daily average $PM_{2.5}$ are shown in Figure 12, and are very similar to the results for the baseline case shown in Figure 6. Similar to O_3, $PM_{2.5}$ drops significantly between 2007 and 2008, but unlike O_3, $PM_{2.5}$ continues to drop in the following years. As in the baseline case, the years 2009 to 2014 are all significantly lower than 2005 at the $\alpha = 0.001$ level. In addition, the two-fold cross-validation tests (described in Section 2.6 below) show little difference in the observed trend, with both randomly-distributed halves of the dataset showing slight increases in 2006 and 2007 followed by dramatic decreases.

The GAM evaluation plots in Figure 13 also indicate a poorer fit for $PM_{2.5}$ than for O_3, as the residuals show a long positive tail and the variance of the residuals is a strong function of the value of the linear predictor, as was the case for the baseline GAMs.

Table 12 lists the meteorological predictors in each urban area that were not significant at the α=0.001 level for maximum MDA8 O_3 and maximum daily average $PM_{2.5}$. Note that solar radiation measurements were not available for San Antonio or Austin/Round Rock.

In addition, we examined the smooth functions fit for each predictor for similarities and differences between the urban areas. For maximum MDA8 O_3:

- The afternoon mean temperature functions all show increasing O_3 with increasing temperature, but the fits for DFW, SA, and ARR flatten out for temperatures > 30 ^oC, while the other areas show no such flattening off.
- O_3 generally increases with increasing diurnal temperature change, but the effect is weak.
- The daily average RH functions all show decreasing O_3 with increasing RH, but the effect is relatively weak in HGB.
- O_3 generally increases with dew point temperature up until 10-15 ^oC, after which point O_3 decreases. This is consistent with the competing effects of temperature and humidity on O_3 production.
- O_3 decreases with daily average wind speed for all urban areas, but the effect is strongest in HGB and SA.
- All urban areas except BPA show increasing O_3 with increasing stability (*T_diff_850mb*). However, O_3 decreases at the highest values of *T_diff_850mb* for SA (-5 to 0 ^oC)
- Daily wind direction generally has little impact on the O_3, and is likely just fitting noise.
- O_3 decreases with HYSPLIT back-trajectory distance up to ~500 km, at which point it becomes highly uncertain due to the low number of points, but may begin to increase.
- All the urban areas show a drop in O_3 at a HYSPLIT bask-trajectory bearing of ~150^o (southeast), likely due to reduced background O_3 from flows from the Gulf of Mexico.
- The day-of-year function shows a slight decrease over the length of the O_3 season for all urban areas, with an area of nearly flat slope at ~200-225 Julian days (July-August).

For maximum daily average $PM_{2.5}$:

- All urban areas generally show $PM_{2.5}$ increasing with afternoon mean temperature, but the effect is fairly weak for ARR, and SA, DFW, and HGB suggest that the trend flattens out or reverses at temperatures > ~30 ^oC.
- The fits for average RH generally peak at 60-70% and fall off at lower and higher RH values, although SA and ARR show a second peak at the lowest extreme values (~20%).
- $PM_{2.5}$ generally increases with increasing temperature at 925 mbar, but HGB shows a significant increase at the lower extreme.
- $PM_{2.5}$ generally trends down with increasing daily average wind speed, but HGB and BPA show an upward trend for wind speeds greater than 6 m/s, possibly related to marine aerosol production.
- All urban areas show increasing $PM_{2.5}$ with increasing stability (*T_diff_850mb).*
- $PM_{2.5}$ decreases with HYSPLIT back-trajectory distance up to ~500 km, at which point it becomes flatter and highly uncertain due to the low number of points. The DFW fit is fairly flat, showing little dependence on back-trajectory distance.
- All urban areas show a maximum for $PM_{2.5}$ around a HYSPLIT bask-trajectory bearing of ~60^o (northeast). DFW, SA, ARR, and TLM show a minimum around 320^o (northwest), possibly due to the relative difference in the $PM_{2.5}$ concentrations in the western and eastern US. However, the urban areas near the Gulf of Mexico (HGB and BPA) have a minimum around ~150^o (southeast), likely due to flows from the Gulf of Mexico.

- $PM_{2.5}$ generally decreases with increasing solar radiation, possibly due to increased cloudiness leading to more rapid oxidation of SO_2 into aerosol sulfate.
- The day-of-year functions for all urban areas are lower in the summer, likely reflecting the higher mixing heights in this season. The maximum is generally between 50-100 Julian days (around March), and ARR and SA show a secondary maximum at ~200 Julian days (July).

We also compared the functional forms in the extended GAMs to those of the baseline GAMS described in Section 2.4. For O_3, although the exact predictors used varied between the models, the functional shapes for temperature, RH, stability, and HYSPLIT 24-hour back-trajectory bearing and distance were very similar between the two models. However, the shape of the day-of-year function changed dramatically, and the daily wind speed dependence in the extended GAMS was generally stronger than the afternoon and morning wind speed effects in the baseline GAMs. For $PM_{2.5}$, the functional shapes for temperature, RH, stability, wind speed, HYSPLIT 24-hour back-trajectory bearing and distance, and day-of-year were all very similar between the baseline and extended models.

Table 12. Meteorological predictors that were not significant at the α=0.001 level for the small extended GAMs (gam03_extended).

Urban Area	MDA8 O_3	Daily-Average $PM_{2.5}$
DFW	*None*	*HYSPLIT_DIST..m.*
HGB	*NCDC.Avg.RH*	*None*
SA	*None*	*SolarRadiation.Langy.min. (not measured)*
ARR	*None*	*SolarRadiation.Langy.min. (not measured)*
BPA	*T_diff_850mb (dropped)*	*T_925mb*
TLM	*None*	*None*

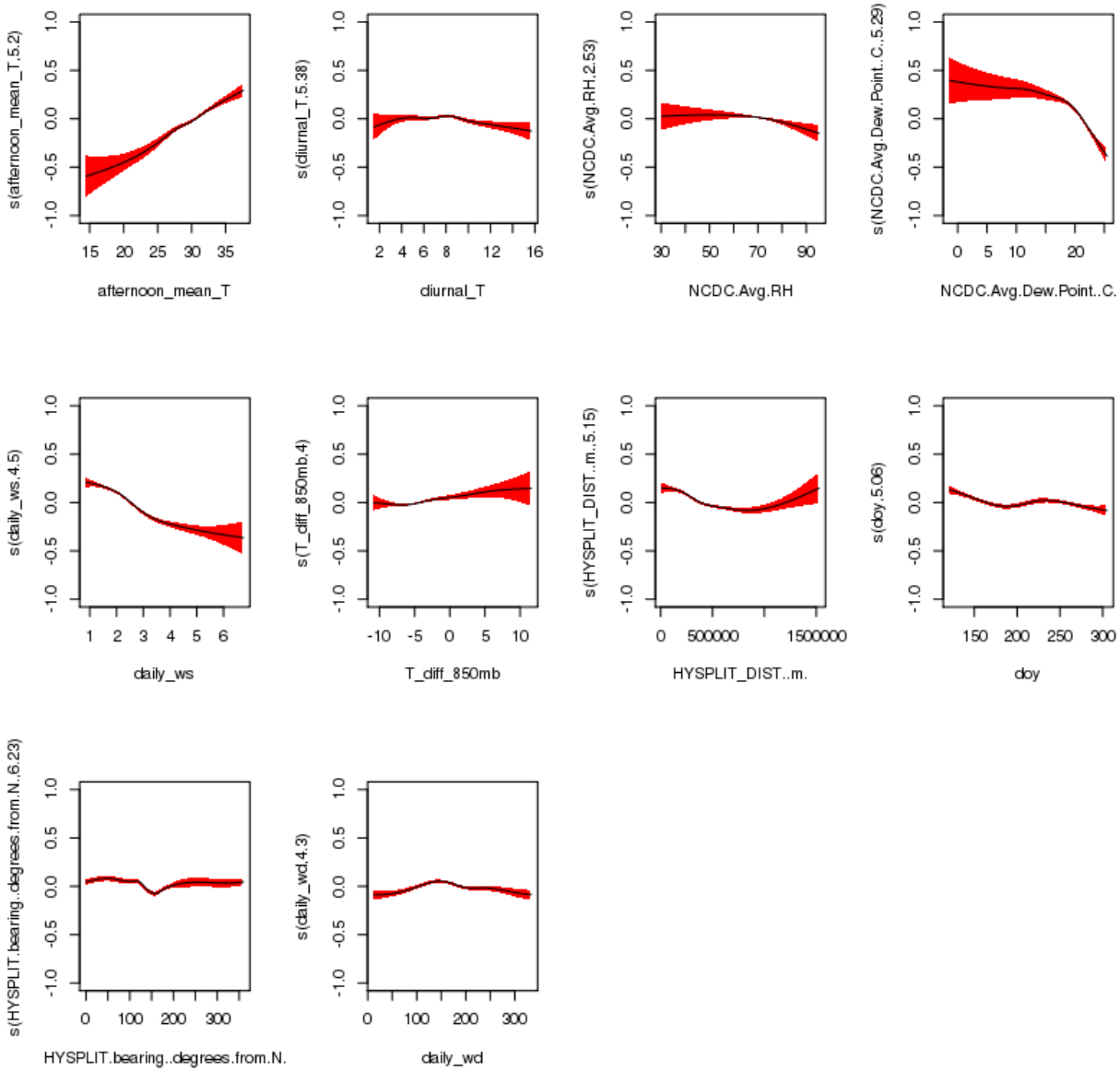

Figure 8. Smooth functions for the small extended GAM (gam03_extended) fit to HGB MDA8 O_3 data.

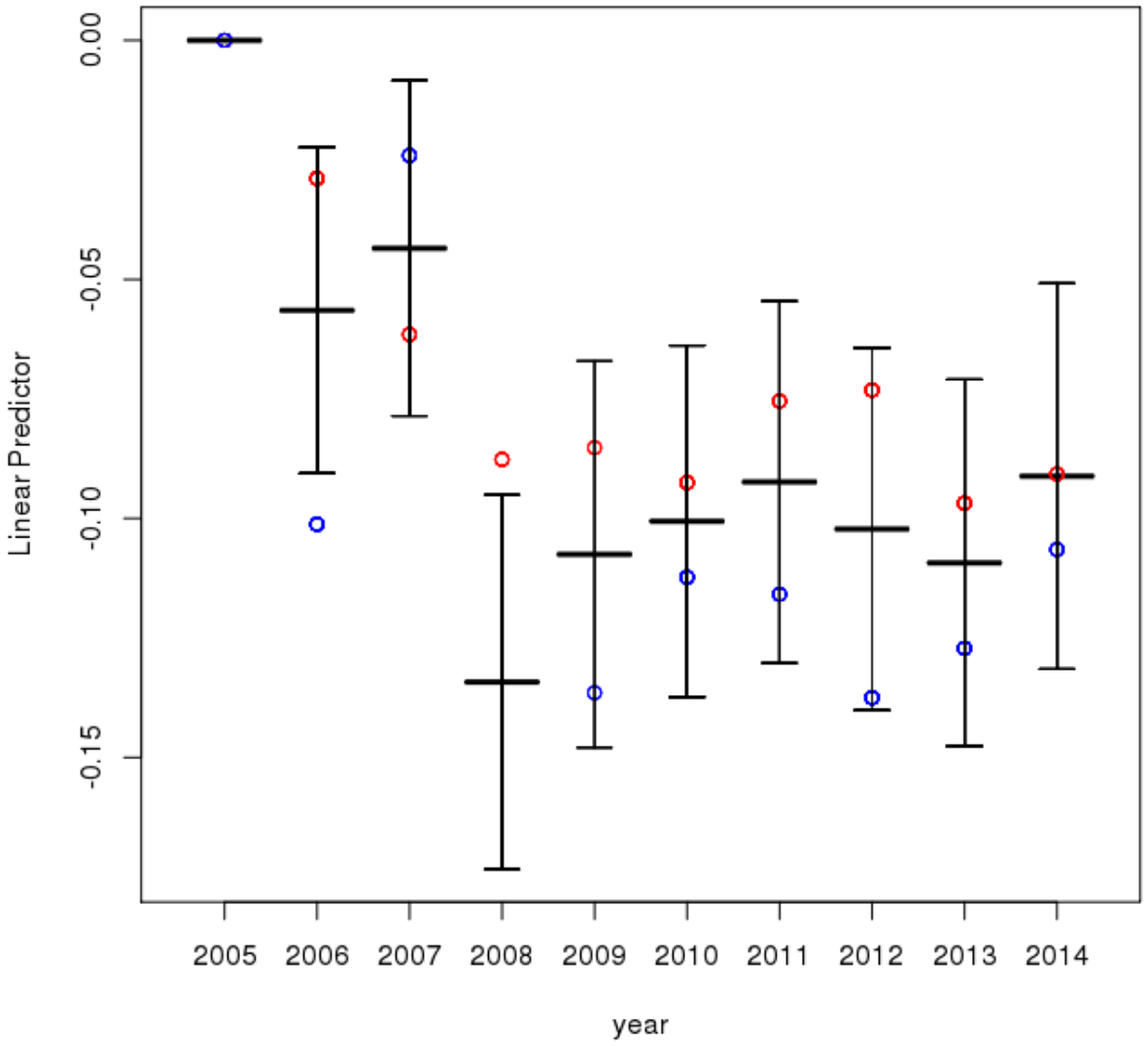

Figure 9. Year-to-year deviations from 2005 for the small extended GAM (gam03_extended) fit to HGB MDA8 O_3 data.

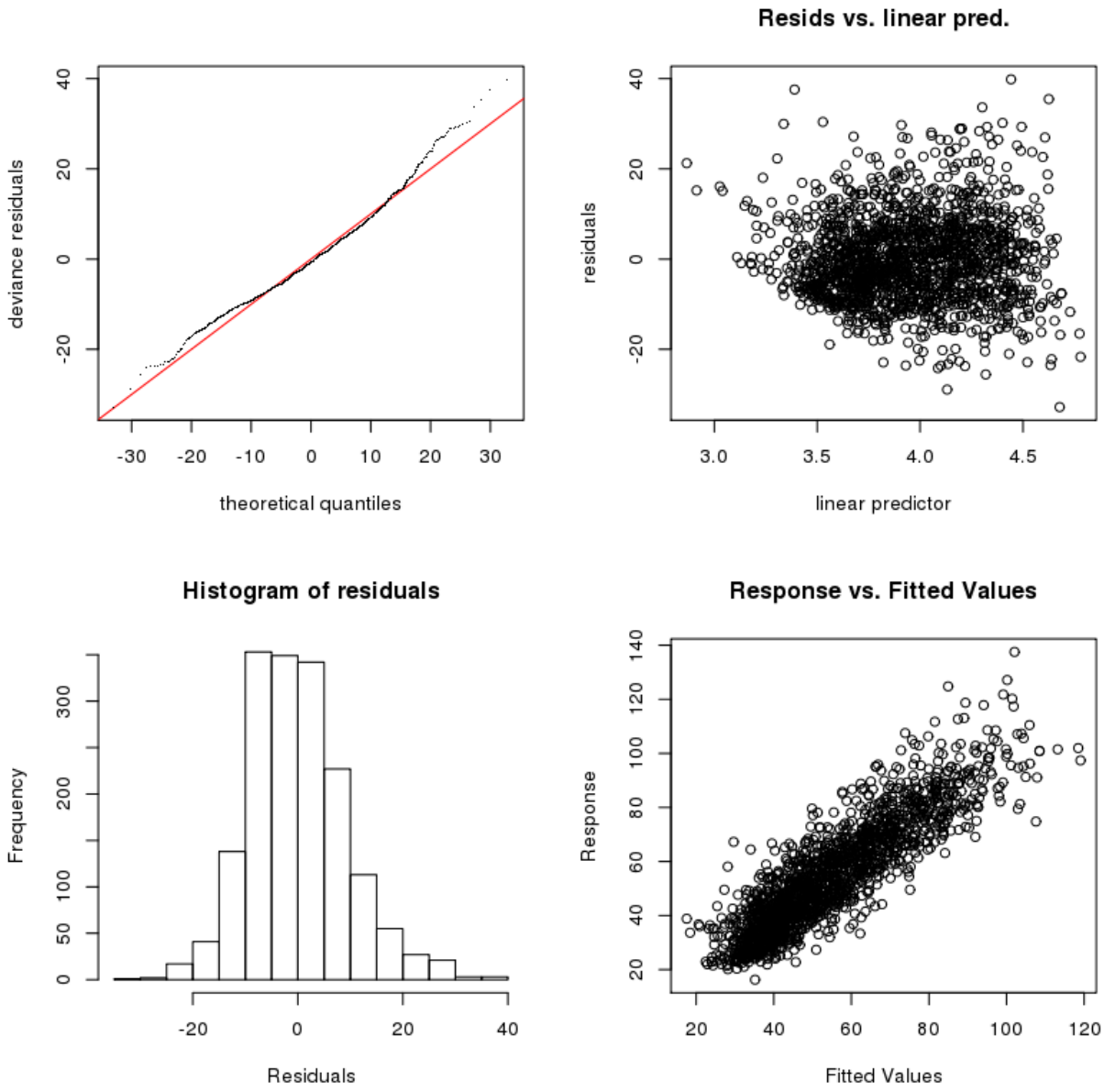

Figure 10. GAM evaluation plots for the small extended GAM (gam03_extended) fit to HGB MDA8 O_3 data.

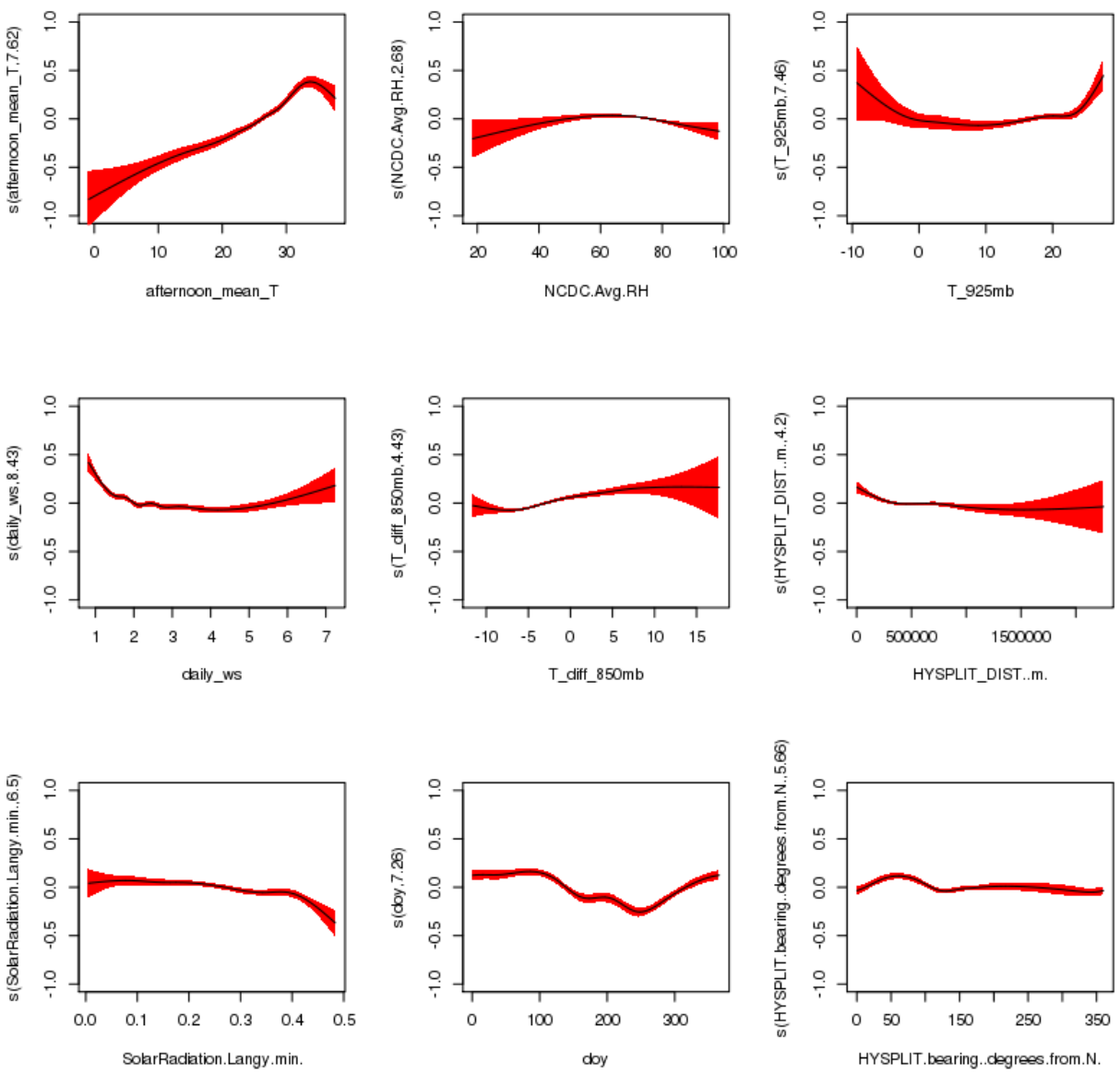

Figure 11. Smooth functions for the small extended GAM (gam03_extended) fit to HGB daily average $PM_{2.5}$ data.

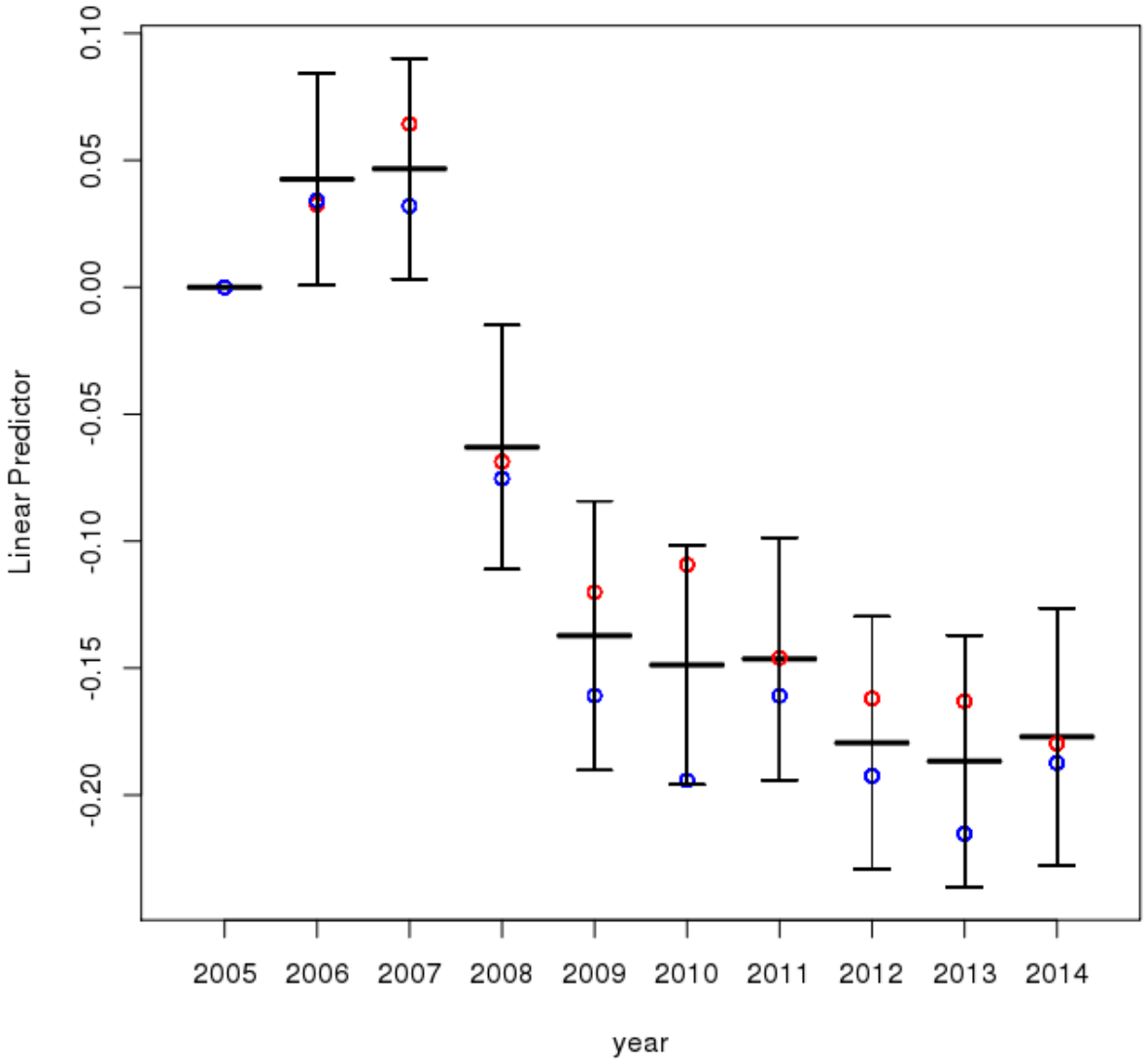

Figure 12. Year-to-year deviations from 2005 for the small extended GAM (gam03_extended) fit to HGB daily average $PM_{2.5}$ data.

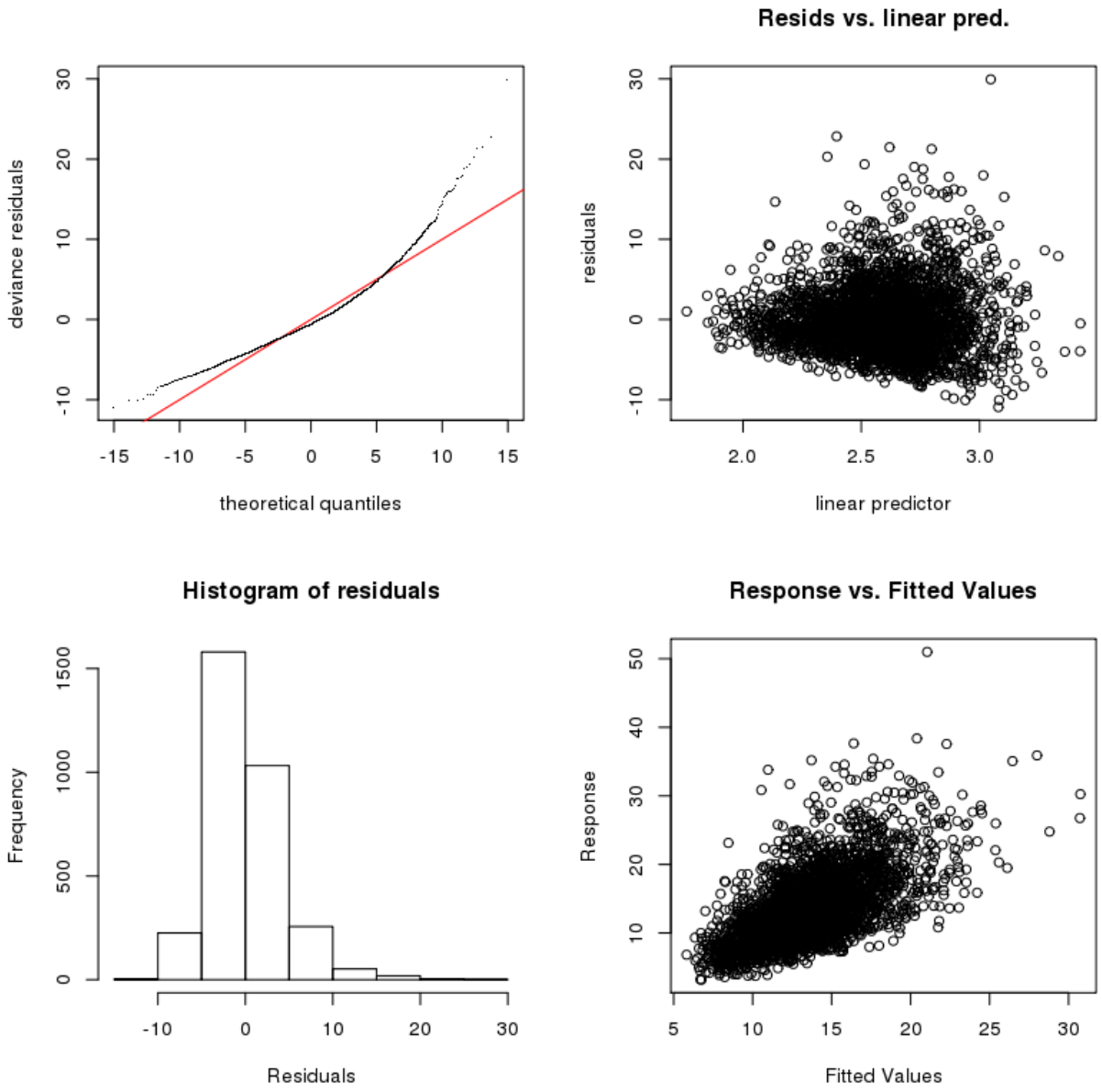

Figure 13. GAM evaluation plots for the small extended GAM (gam03_extended) fit to HGB daily average $PM_{2.5}$ data.

2.6 Cross-Validation Analysis

In order to test for over-fitting in our GAMs, as well as to test the robustness of our results for the functional relationships between the meteorological predictors and O_3 and $PM_{2.5}$, we performed a two-fold cross-validation experiment for each GAM. To do this, the original dataset was randomly separated into two halves (data sets 1 and 2). We then fit two GAMs (hereafter m_1 and m_2) using the two halves of the data. The performance of these GAMs on the half of the data they were not trained on was then compared to the performance of the corresponding GAM that was fit on all the data (hereafter m_{tot}).

Figure 14 shows scatterplots of the GAM-predicted (x-axis) versus the measured (y-axis) values of maximum daily average $PM_{2.5}$ for the Houston/Galveston/Brazoria area using gam03_extended. We can see that the performance of m_1 and m_2 on their respective test data sets is similar to the performance of the original GAM m_{tot}. This can also be seen in Table 13, which shows the root-mean-square (RMS) differences between the GAM-predicted and measured O_3 and $PM_{2.5}$ values for gam03_extended. The change in the RMS between m_{tot} and m_1 and m_2 is generally small (< 1 ppbv for O_3 and < 0.25 ug m^{-3} for $PM_{2.5}$). As the training set and testing set RMS errors are thus similar, we conclude there is little evidence of overfitting in our GAMs.

However, the individual functional forms relating the meteorological and date predictors to O_3 and $PM_{2.5}$ can occasionally be significantly different between m_{tot}, m_1, and m_2, suggesting that these relationships, although statistically significant, may not be robust or scientifically meaningful. For example, Figure 15 shows the Houston fits for maximum daily average $PM_{2.5}$ versus HYSPLIT back-trajectory bearing for m_{tot} (black with error bars), m_1 (red), and m_2 (blue) for gam03_extended. Predicted values for 200 randomly selected data points are plotted. We see m_2 significantly differs from m_{tot} between 0^o and 100^o, suggesting the functional form from m_{tot} may not be robust in this region. Plots similar to Figure 15 for all GAMs and their terms are contained in the deliverable, as described in Section 3.6. Other "suspicious" functional forms for $PM_{2.5}$ and O_3 in the gam03_extended fits are listed in Table 14, but we note that as these are for a single random division of the dataset, these results merely indicate a potential problem, but do not by themselves prove that the functional relationships are incorrect.

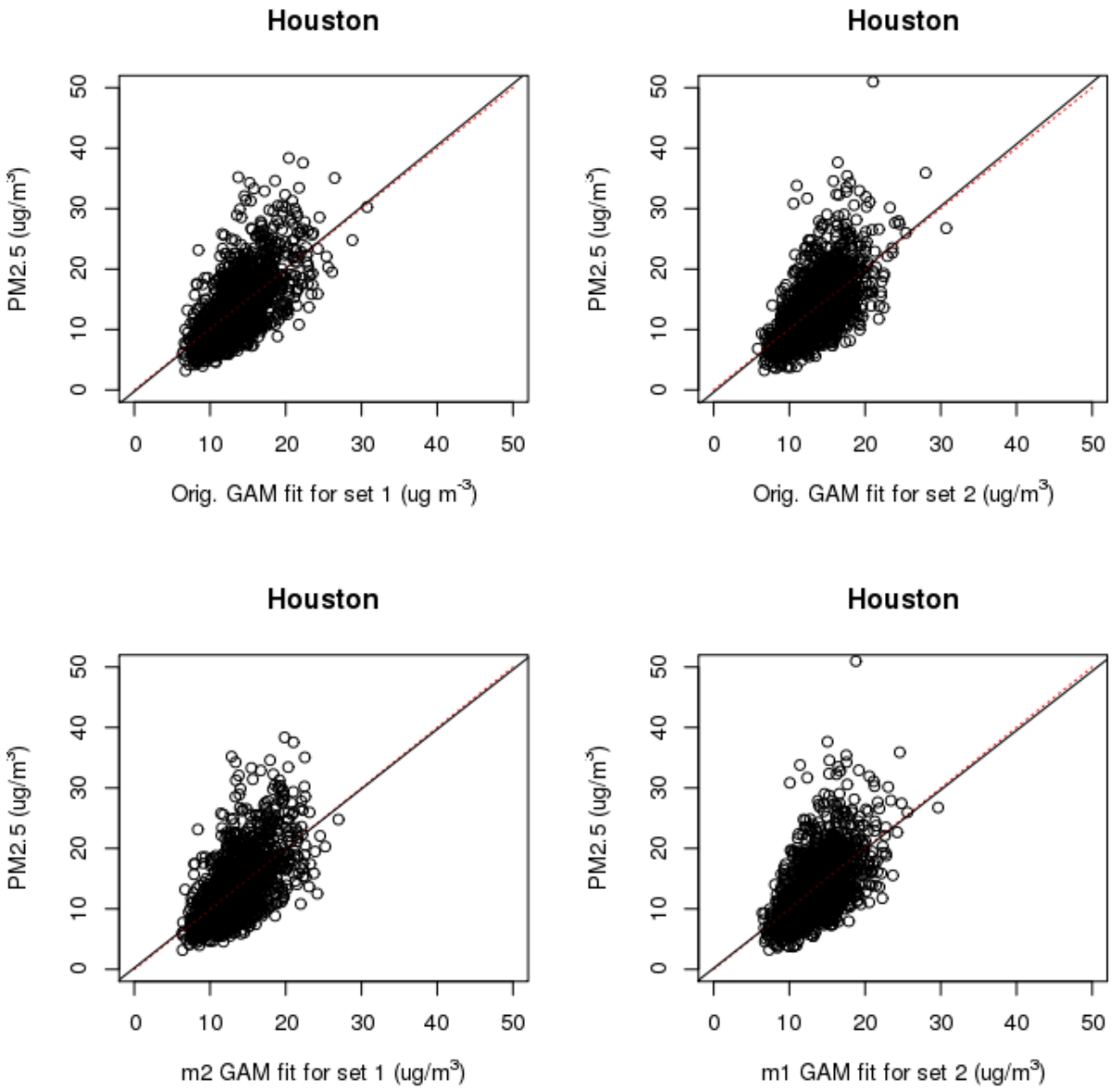

Figure 14. Scatterplots for the GAM-predicted (x-axis) versus the measured (y-axis) values of maximum daily average $PM_{2.5}$ for the Houston/Galveston/Brazoria area using gam03_extended. The top row uses m_{tot} to predict the first (left) and second (right) of the randomly distributed halves of the dataset. The bottom row uses m_2, which was trained on data set 2, to predict the "test" data set 1 (left) and uses m_1 to predict data set 2 (right). The black line is a linear fit of the predicted to actual values, while the red dashed line is the 1:1 line.

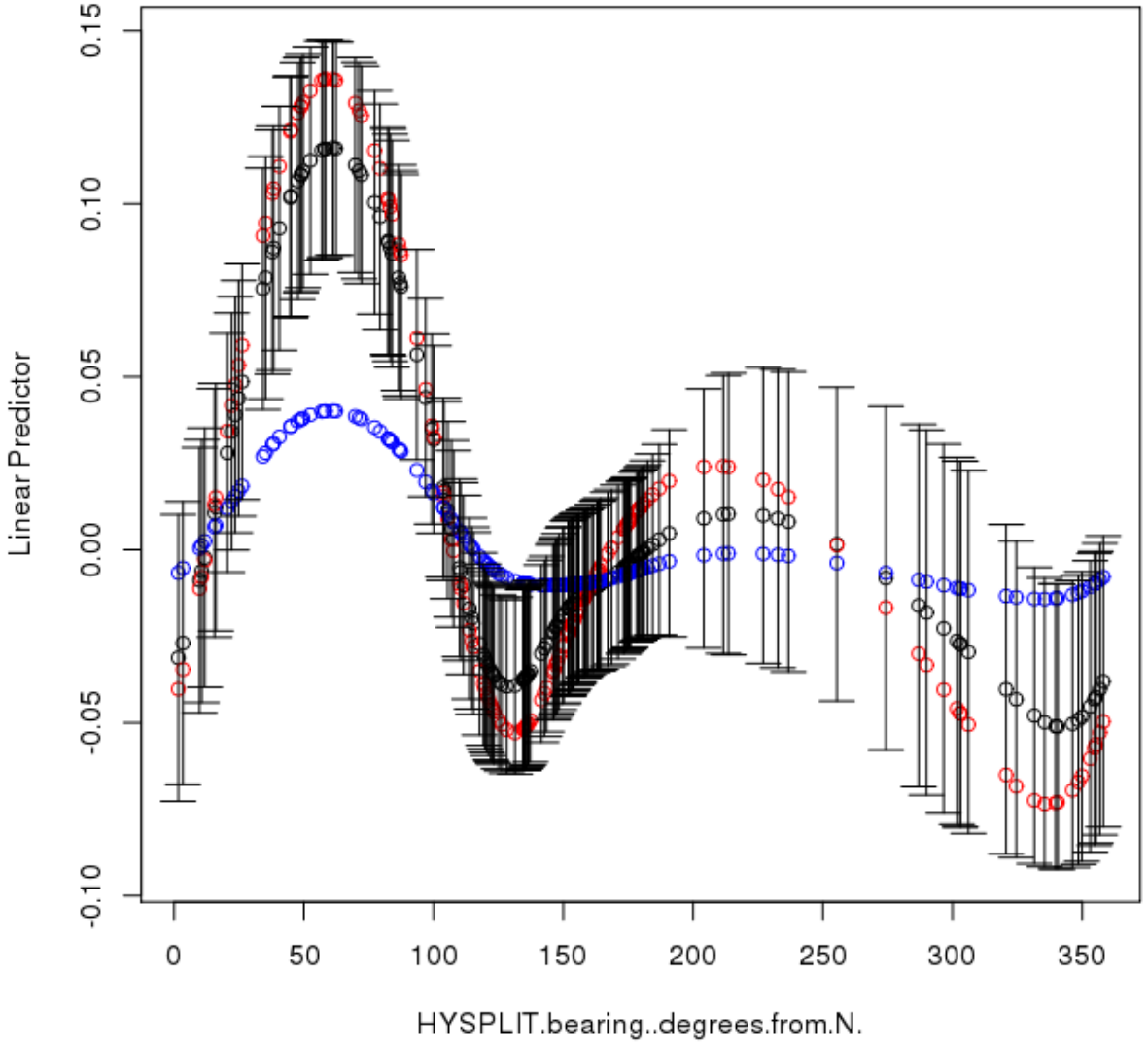

Figure 15. Houston/Galveston/Brazoria fits for maximum daily average $PM_{2.5}$ versus HYSPLIT back trajectory bearing for m_{tot} (black with error bars), m_1 (red) and m_2 (blue) for gam03_extended. Predicted values for 200 randomly selected datapoints are plotted.

Table 13. Cross-validation root-mean-square (RMS) results for gam03_extended.

Urban Area	MDA8 O_3 (ppbv)				Daily Average $PM_{2.5}$ (μg m^{-3})			
	Data Set 1		Data Set 2		Data Set 1		Data Set 2	
	m_{tot}	m_2	m_{tot}	m_1	m_{tot}	m_2	m_{tot}	m_1
DFW	7.79	8.27	8.13	8.56	3.95	4.07	3.90	4.03
HGB	9.09	10.07	9.70	10.53	4.08	4.26	4.15	4.27
SA	7.37	7.94	7.20	7.76	3.77	3.94	3.95	4.07
ARR	7.04	7.67	7.23	7.72	3.79	3.93	3.79	3.89
BPA	8.35	9.11	8.70	9.21	4.80	5.02	4.71	4.93
TLM	7.80	8.14	7.46	7.76	4.45	4.56	3.41	3.55

Table 14. "Suspicious" fits that show significantly different functional forms between m_{tot}, m_1, and m_2 for gam03_extended.

Urban Area	MDA8 O_3	Daily Average $PM_{2.5}$
DFW	*HYSPLIT.bearing..degrees.from.N., diurnal_T*	*NCDC.Avg.RH*
HGB	*T_diff_850mb*	*HYSPLIT.bearing..degrees.from.N., SolarRadiation.Langy.min*
SA	None	None
ARR	None	None
BPA	None	*HYSPLIT.bearing..degrees.from.N., NCDC.Avg.RH*
TLM	None	*HYSPLIT_DIST..m., T_diff_850mb*

3 File Descriptions

This section describes all of the files included in the deliverable. Figure 16 is a flow chart showing the processing from the initial data sources to the final CSV file used as input for the GAM fitting. These files are described in Sections 3.1 to 3.4. Figure 17 shows the scripts that use the CSV file produced at the end of Figure 16 to produce and evaluate the GAMs. These scripts and the output files produced are described in Sections 3.5 and 3.6, respectively.

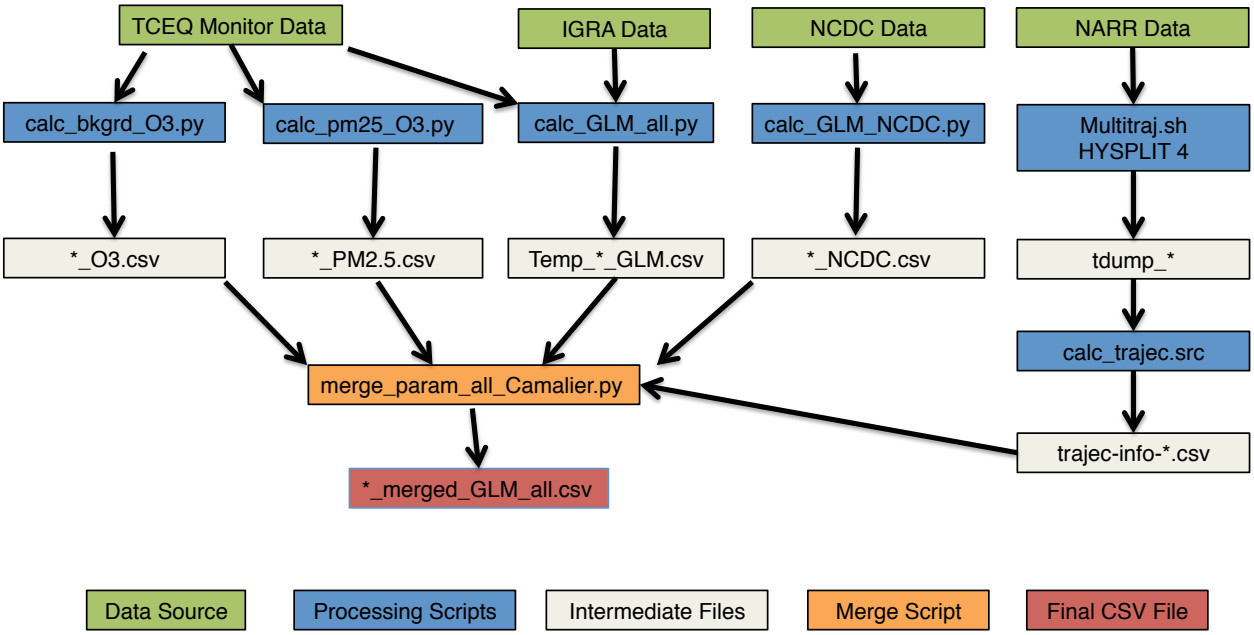

Figure 16. Flow chart showing the processing from the original data sources (green boxes) to the final CSV file (red box) that is used as input for the GAM fitting scripts.

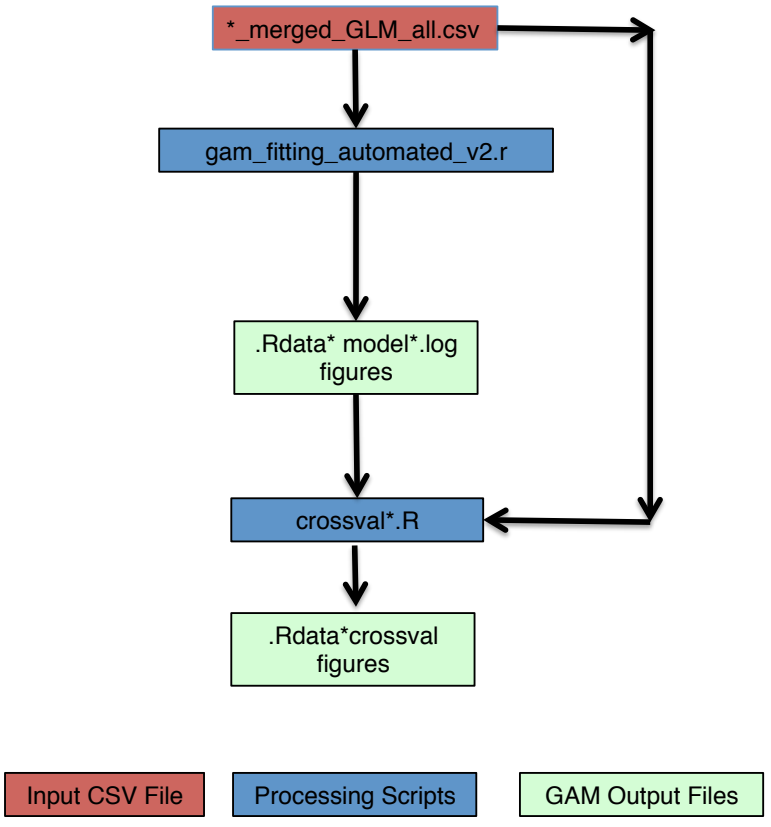

Figure 17. Flow chart showing the processing from the input CSV file generated at the end of Figure 16 (red box) to the GAM output files (light green box).

Note that all R scripts below were run using R version 3.1.1 (2014-07-10) and package mgcv v1.8-0 on an x86_64-redhat-linux-gnu (64-bit) platform (CentOS release 5.11 Final) with Dual-Core AMD Opteron™ Processor 2218 and 8 GB RAM per core. All python scripts were run using Python v3.4.3 and Ipython v3.1.0 on a MacBook Pro with a 3.1 Ghz Intel Core I7 processor and 16 GB of RAM running Mac OS X Yosemite Version 10.10.3. The HYSPLIT runs were performed using a K shell (ksh) script on a Linux cluster running SUSE Linux Enterprise Server v11 (x86_64) with 12 Intel® Xeon® CPUs (X5650 @ 2.67GHz) and 4 GB RAM per processor. The Microsoft Excel spreadsheets were made using Microsoft Excel for Mac v14.5.2. All scripts should run on any Linux or Mac OS X system with the correct versions of R, Python, and Microsoft Excel installed.

3.1 Input data *(./data/)*

This directory contains the raw IGRA and NCDC data used in this project. The raw TEMIS monitor data provided by TCEQ is not included in the deliverable.

3.1.1 IGRA Data (*./data/IGRA_data/*)

The Integrated Global Radiosonde Archive (IGRA) provided upper atmosphere data used to derive some of the meteorological predictors. The sites selected are described in Table 3, with the data files named *#####.dat* according to the ID number of the selected sites along with a *readme.txt* file that describes the data format and measurements. The relevant measurements include the geopotential height, temperature and dewpoint depression at several altitudes with -8888 values indicating original value has been removed by IGRA and -9999 was never present.

3.1.2 NCDC data (*./data/NCDC_data/)*

This directory contains the National Climatic Data Center (NCDC) Integrated Surface Hourly (ISH) dataset used to get estimates of surface pressure and relative humidity, as this data was not generally available in the TCEQ dataset. Each urban area has a directory within *./data/NCDC_data/* that contains the raw data (*###dat.txt*) and two station description files (*###inv.txt* and *###stn.txt*). The data from the station description files is also in Table 4. The raw data file contains daily data from 2005-2014. Missing data is indicated by ***.

3.2 Data Processing Scripts (*./scripts/*)

- *./scripts/calc_bkgrd_ozone.py* : This script reads in the ozone monitor data provided by the TCEQ to calculate the maximum daily 8 hour average (MDA8) O_3 for each urban area. After filtering out non-data the script derives the maximum and minimum MDA8 for all urban locations, as well as the minimum (background) MDA8 O_3 value for all selected background sites according to the technique described in Section 2.1.1. The selected background sites are listed in Table 15 as well as in the script itself. The produced CSV files are input to the script *./scripts/merge_param_all_Camalier.py* (described below), which will combine O_3 daily values with the GLM parameters. The outputs of this script were previously supplied to TCEQ as Deliverable 3.1.
- *./scripts/calc_pm25.py* : This script reads in the $PM_{2.5}$ monitor data provided by the TCEQ to calculate the maximum and minimum daily $PM_{2.5}$ concentrations for all urban locations, as well as the daily minimum (background) concentrations for the selected background sites according to the technique described in Section 2.1.1. These background sites are listed in Table 15 as well as in the script itself. The produced CSV files are input to the script *./scripts/merge_param_all_Camalier.py* (described below),

which will combine $PM_{2.5}$ daily values with the GLM parameters. The outputs of this script were previously supplied to TCEQ as Deliverable 3.1.

- *./scripts/calc_GLM_all.py* : This script reads in TCEQ monitor site and IGRA (upper atmosphere) measurements to derive daily GLM parameters described in Appendix A and in the script itself. It performs all the necessary conversions (ex. Fahrenheit to Celsius, mph to m s^{-1}) and derivations (ex. wind direction u component, dewpoint to RH based on August-Roche-Magnus approximation), to compile the full list of daily meteorological predictors, except those from the NCDC (described below). See the script for full details on all conversions and derivations. It creates the intermediate files for each urban area located in *./csv_files/intermed_files/TCEQ_files/*, and these output files are used in *./scripts/merge_param_all_Camalier.py*.
- *./scripts/calc_GLM_NCDC.py* : This script reads in the NCDC data to derive daily meteorological predictors described in Appendix A and indicated as an NCDC parameter. It performs all the necessary conversions (ex. Fahrenheit to Celsius) and derivations (ex. Apparent Temperature according to the National Digital Forecast Database). See the script for full details on all conversions and derivations. It creates the intermediate files for each urban area located in */csv_files/intermed_files/NCDC_files/,* and these output files are used in *./scripts/merge_param_all_Camalier.py*.
- *./scripts/merge_param_all_Camalier.py* : This script reads in all intermediate files described above for each urban location. This includes the daily maximum and background concentrations for O_3 and $PM_{2.5}$, as well as daily values for all meteorological predictors listed in Appendix A. It aligns the date for all files, checks for missing data and replaces with 'nan' if there is no data. It creates the final merged files that are located in */csv_files/final_files* and are used in the GAM fitting scripts described in Section 3.5.

Table 15. AQS site numbers for the selected background sites for each urban area.

Urban Area	Total # of Sites	# of Background Sites	AQS Site Numbers of Background Sites
DFW	28	11	481210034, 481211032, 481215008, 481391044, 482210001, 482311006, 482510003, 482570005, 483491051, 483670081, 484390075
HGB	69	31	480390618, 480390619, 480391003, 480391004, 480391016, 480710013, 481570696, 481670697, 481671034, 481675005, 482010029, 482010066, 482010552, 482010553, 482010554, 482010555, 482010556, 482010557, 482010558, 482010559, 482010560, 482010561, 482010563, 482010617, 482011042, 482011050, 482910699, 483390078, 483390698, 483395006, 483739991
SA	15	8	480290059, 480290501, 480290502, 480910503, 480910505, 481870504, 481870506, 481875004
ARR	12	8	482090675, 480210684, 481490001, 482090614, 482091675, 484530020, 484910690, 484916602
BPA	17	5	482450022, 482450101, 482450628, 483611001, 483611100
TLM	4	4	484230007, 481830001, 482030002, 480370004

3.3 HYSPLIT

3.3.1 HYSPLIT run script (*./HYSPLIT_runs_out/*)

- *./HYSPLIT_runs_out/multitraj.sh* : A K shell script that runs the 24-hour HYSPLIT 4 back-trajectories for each urban region for the 2005-2014 period described in Section 2.2. The script consists of multiple nested loops over inner to outer city, day, month and year. Each time through the loop the city, day, month, and year information is written to the CONTROL text file that is input to HYSPLIT and the HYSPLIT run is executed. Upon run completion the trajectory endpoint is extracted from the trajectory output file, tdump, and appended to the appropriate tdump_*city* CSV file.

3.3.2 HYSPLIT back trajectory endpoints (*./HYSPLIT_runs_out/*)

- *./HYSPLIT_runs_out/tdump_** : One of six intermediate CVS files generated from the *./HYSPLIT_runs_out/multitraj.sh* script, one for each urban area of interest. * is a 3-letter code indicating the urban area. The first line in each file lists the 3-letter city code and the latitude and longitude of the trajectory origin. The starting back trajectory elevation is always 300 m above ground level (agl) and not included in

these files. The rest of the lines are the endpoint time and location data, one line per endpoint. The lines include the following:

- Trajectory run - will always be 1 in this application, ignore
- Trajectory number – will always be 1 in this applications, ignore
- YEAR – 2-digit format
- Month
- Day
- Hour – always 18 UTC
- Minute – always 0
- Second –always 0
- Trajectory age – always -24 (indicating a 24 hour back trajectory)
- Latitude
- Longitude- west is negative
- Elevation- meters AGL
- Pressure – hPa

3.3.3 HYSPLIT distance and bearing calculation script and output (*./hysplit_trajec/)*

- *./hysplit_trajec/calc_trajec.src* : This R script takes the 24 hour back-trajectory endpoint files from the *./HYSPLIT_runs_out/* directory and calculates the distance and bearing from the starting point to the end point of the trajectory using the R functions *bearing* and *distMeeus* from the *geosphere* package as described in Section 2.2.
- *./hysplit_trajec/trajec-info-*.csv* : CSV file produced by *./hysplit_trajec/calc_trajec.src* that contains the distance and bearing for the back trajectories. A separate file exists for each urban area. These files are used as inputs by *./scripts/merge_param_all_Camalier.py* (Section 3.2).

3.4 Processed Input Data Files in CSV Format (*./csv_files/*)

3.4.1 Intermediate CSV Files (*./csv_files/NCDC_files/ and ./csv_files/TCEQ_files/*)

These files include the meteorological predictors derived from the NCDC, TCEQ and IGRA datasets described in Section 3.1 using the scripts described in Section 3.2 (*./scripts/calc_GLM_NCDC.py* and *./scripts/calc_GLM_all.py* respectively). They contain daily GLM values (Appendix A) for all urban locations from 2005-2014 and are used as input by *./scripts/merge_param_all_Camalier.py*.

3.4.2 Final CSV Files (*./csv_files/final_files/*)

These files are created by *./scripts/merge_param_all_Camalier.py* (Section 3.2), which combines all daily meteorological predictors with the O_3 and $PM_{2.5}$ concentrations for each location. The entire file is described in Appendix A and includes daily values from 2005-2014, with missing values indicated by 'nan'. These files are used as inputs by the GAM scripts described in Section 3.5.

3.5 GAM scripts (*./full_gam_fits/*)

3.5.1 Correlation Screening

- *./full_gam_fits/cor_test_mja.R* : A log of R commands that shows how to read in the final CSV data files and assess a set of variables for correlation, as described in

Section 2.5.1. Note that this is NOT a script you can run as-is, it merely is a record of the necessary commands.

- *./full_gam_fits/cor_test_results_ozone.xlsx* : A Microsoft Excel spreadsheet showing the families of variables tested in the initial correlation screening and the selected variables for ozone in each city.
- *./full_gam_fits/cor_test_results_pm2.5.xlsx* : Same as above but for $PM_{2.5}$.

3.5.2 GAM Fitting

- *./full_gam_fits/gam_fitting_automated_v2.r* : The main GAM fitting script. The options are described at the top of the script. It takes a CSV data file and arrays specifying types of modeled variables, fits a GAM model (as specified or finds the best fit by eliminating variables), and produces (see Section 3.6):
  - A log of final model diagnostics: summary, gam.check, & table summarizing iterations (if find.best.fit is TRUE). Log may optionally include model summaries for every model iteration (if verbose is TRUE and find.best.fit is TRUE)
  - gam.check plot
  - smooth variable function plots (if create.plots is TRUE)
  - R data object containing final model (mod) and associated variable arrays (factor.vars, linear.vars, cr.vars, and cc.vars). This can be loaded and reused for plots or other diagnostics later in R.
- *./full_gam_fits/automate_gam_fitting.src* : A driver script for *./full_gam_fits/gam_fitting_automated_v2.r* that sets the necessary inputs.

3.5.3 Cross-Validation

- *./full_gam_fits/crossval_pm.R* : An R script that performs a cross-validation check on our $PM_{2.5}$ GAMs. It randomly divides the original dataset into two halves, then fits a GAM to each half separately. The performance of these GAMs on the half of the data they were not trained on is then compared to the performance of the corresponding GAM fit on all the data. The smooth functional fits for all three GAMs are also plotted to check for differences between the two halves. At the top of the script, change "city" and "model" to test the appropriate GAM.
- *./full_gam_fits/crossval_o3.R* : Same as above, but for the O_3 GAMS.

3.6 GAM Output Files (*./full_gam_fits/o3_model/* and *./full_gam_fits/pm2.5_model/*)

The output directories *./full_gam_fits/o3_model/* and *./full_gam_fits/pm2.5_model/* both contain one subdirectory for each urban area (e.g., *./full_gam_fits/o3_model/Houston/*). Each of these urban area subdirectories contains a subdirectory for each of the three GAMs contained in the deliverable, such as:

- *./full_gam_fits/o3_model/Houston/o3gam01_baseline/*
- *./full_gam_fits/o3_model/Houston/o3gam02_extended/*
- *./full_gam_fits/o3_model/Houston/o3gam03_extended/*

The files contained in each of these model directories are described below, using the file names from *./full_gam_fits/o3_model/Houston/o3gam03_extended/* as an example :

- *.RData_o3gam03_extended_Houston* : An R data file containing the GAM as an element in the list 'mod' (e.g., for this case it the GAM can be accessed as mod[['o3gam03_extended']]). The script *./full_gam_fits/crossval_pm.R* shows an example of how to load the GAM object (L32-35) and rebuild the GAM formula (L37-45) using this data file.
- *model_results_Houston_20150626.log* : The log file for the GAM fit as produced by the script *./full_gam_fits/gam_fitting_automated_v2.r*. The first line shows the input data file from *./csv_files/final_files/*. The summary of the final selected GAM (after any automated dropping of variables) is in this file after the phrase "FINAL MODEL DIAGNOSTICS". A table at the end of the file summarizes the variables that were tested and dropped by the automated selection procedure described in Section 2.3.
- *plot_o3gam03_extended_Houston_smoothfunc-noresid.png* : A figure showing the smooth functional fits for the GAM, as in Figure 2.
- *plot_o3gam03_extended_Houston_smoothfunc.png* : As above, but with the partial residuals overplotted.
- *gam.check_o3gam03_extended_Houston.png* : A figure showing the standard diagnostic plots for the GAM, as in Figure 4.
- *cross_val/* : A subdirectory containing the output of the cross-validation scripts *./full_gam_fits/crossval*.R.* These files include:
    - *.RData_o3gam03_extended_Houston_crossval* : An R data file containing the original GAM fit (mtot) and the two fits to the randomly selected halves of the data (m1 and m2). The seed number (seed.num) used in the cross-validation script is also stored, as are the indices of the halves of the data used to fit m1 and m2 (ind1 and ind2) and the indices of the 200 randomly-selected data points used to make the cross-validation figures (ind3).
    - *crossval_scatter_Houston_o3gam03_extended.png* : Scatter plots of the predicted (x-axis) versus actual (y-axis) MDA8 O_3 or daily average $PM_{2.5}$ values, as in Figure 14.
    - *cross val_m1_Houston_o3gam03_extended.png* : A figure showing the smooth functional fits for the GAM m1 fit to the data in ind1, similar to Figure 2.
    - *cross val_m2_Houston_o3gam03_extended.png* : Same as above but for the GAM m2 fit to the data in ind2.
    - *crossval_terms*.png* : Plots of the smooth function predictions for 200 randomly selected data points (ind3), similar to Figure 15. The files contain the column names of the variables used in the fit. The y-axis scale is the scale of the "linear predictor", i.e. the deviation of the natural logarithm of the MDA8 O_3 or the daily average $PM_{2.5}$ in $\mu g\ m^{-3}$ from its mean value. The black center bar is the mean value while the error bars are the 95% confidence intervals. The red and blue circles are the mean values from the two-fold cross-validation analysis of Section 2.6.

4 Quality Assurance Steps

In addition to the analyses described in Section 2, other quality assurance checks were made. All scripts used in this project were inspected by teams members different from the original

author to ensure they were calculating properly, and any errors noted in early versions were fixed. In addition, if further analysis or feedback from TCEQ uncovers any errors in the provided files, we will correct those and provide TCEQ with corrected files as part of our Final Report.

The project Quality Assurance Project Plan (QAPP) listed several questions that needed to be addressed as part of the GAM evaluation, as well as several required pieces of model documentation. These are addressed below in Sections 4.1 and 4.2, respectively.

4.1 Model Evaluation

The QAPP stated that the evaluation of the GAMs produced in this project would address the following questions:

- *Do the relationships between meteorological variables and O_3 and $PM_{2.5}$ described in the developed GAMs make physical sense given our conceptual models of O_3 and $PM_{2.5}$ emissions, chemistry, and transport?*

  As noted in Sections 2.4.2 and 2.5.2, the functional dependencies in the GAMs between the predictors related to temperature, RH, wind speed, vertical stability, and HYSPLIT bearing are all qualitatively consistent with our conceptual understanding of O_3 and $PM_{2.5}$ emissions, chemistry, and transport.

- *Are these relationships consistent with the scientific literature?*

  As noted in Section 2.4.2, our GAMs for MDA8 O_3 are consistent with those found for eastern US cities by Camalier et al. (2007).

- *Does the change in the relationships between urban areas make physical sense given our conceptual models of O_3 and $PM_{2.5}$ emissions, chemistry, and transport?*

  We find that the general trends of the relationships rarely change significantly between the urban areas. For O_3, the major differences are that DFW, SA, and ARR show the O_3 trend with afternoon temperature flattening out above 30 ^oC and that the impact of relative humidity is fairly weak in HGB. For $PM_{2.5}$, the major differences are between the cities near the Gulf of Mexico (HGB and BPA) and the others, with the cities near the Gulf showing increasing $PM_{2.5}$ at wind speed above 5 m/s and a minimum in $PM_{2.5}$ at a HYSPLIT bearing of 120^o instead of at 320^o.

- *Are the HYSPLIT back-trajectories used in the model development reasonable? How sensitive are these trajectories to the initial location?*

  As noted in Section 2.2, the HYSPLIT back-trajectories used in the model development appear reasonable and generally consistent with the surface wind speed and direction measured near the center of each urban area. The ensemble back-trajectory results suggest that our results are representative of the air masses entering each urban area, but that differences in distance of less than ~100 km and differences in bearing of less than $\sim 20^o$ are unlikely to be significant.

- *How well does the GAM reproduce the testing sets in the cross-validation evaluation?*

  As noted in Section 2.6, the two-fold cross-validation showed that the GAMs fit to half of the data fit the other half of the data nearly as well as the GAMs fit to all of the data.

- *Does the cross-validation evaluation of the models show evidence of over-fitting?*

As noted in Section 2.6, there is no evidence of over-fitting in the overall MDA8 O_3 and daily average $PM_{2.5}$ predictions. However, the functional relationships between the meteorological predictors and O_3 and $PM_{2.5}$ are occasionally sensitive to which half of the dataset is used for the fit, and so caution must be used in interpreting these relationships.

- *Under what conditions are the GAMs expected to be valid? What conditions give exceptionally large residuals?*

  Strictly speaking, the GAMs are only expected to be valid during the periods for which they were fit, and when the data is taken from the sources and sites noted in this memo. Extrapolations to other times and monitoring locations may be problematic, and the GAMs ability in this regard has not been assessed in this project.

  We have not yet identified any set of necessary or sufficient conditions that lead to large residuals in the GAMs. We will continue investigating this and provided updated results with our final report.

4.2 Model Documentation

The QAPP listed several required parts for the model documentation. These are listed below along with where to find the corresponding documentation in this memo.

- *The final model description, hardware and software requirements, including programming language, model portability, memory requirements, required hardware/software for application, and data standards for information storage and retrieval*

  The final descriptions of the GAMs are given in Sections 2.4 and 2.5. The software versions and computers used to run the scripts supplied in the deliverable are documented in the beginning of Section 3.
- *The equations on which the model is based*

  The main GAM equation is given in Section 2.3. More details on the GAM fitting procedure can be found in Wood (2006).
- *The underlying assumptions used in the model development*

  The GAM development procedure and any underlying assumptions are discussed in Section 2. Underlying assumptions of the *mgcv* R package used to perform the fits are discussed in Wood (2006).
- *Flow charts of model inputs, processing, and outputs*

  Figure 16 *and* Figure 17 contain flow charts showing the processing of data from the initial data sources through to the GAMs and their evaluation scripts.
- *Descriptions of the software routines*

  The scripts developed in this project are described in Sections 3.2, 3.3, and 3.5.
- *Data base description*

  The non-TCEQ initial data and the processed intermediate data used to generate the GAMS is contained in the deliverable, as noted in Sections 3.1 and 3.4. The sources of this data are described in Section 2.1.
- *A copy of the source code*

Copies of all scripts developed in this project are contained in the deliverable, as described in Sections 3.2 and 3.5.

- *Explanation of error messages*
  Error messages produced using the GAMs in R are described in the documentation of the *mcgv* package. Error messages in the R and Python scripts supplied in this project are self-explanatory and generally refer to errors in the specified inputs (i.e., missing input files, incorrect parameter settings).
- *Parameter values and sources*
  Parameter values used in the R and python scripts and the sources of those values are documented in the scripts themselves.
- *Restrictions on model application, including assumptions, parameter values and sources, boundary and initial conditions, validation/calibration of the model, output and interpretation of model runs;*
  As noted above, the functional relationships between the meteorological predictors and O_3 and $PM_{2.5}$ are occasionally sensitive to which half of the dataset is used for the fit, and so caution must be used in interpreting these relationships.
- *Limiting conditions on model applications, with details on where the model is or is not suited*
  As noted above, the GAMs are only expected to be valid during the periods for which they were fit, and when the data is taken from the sources and sites noted in this memo. Extrapolations to other times and monitoring locations may be problematic, and the GAMs ability in this regard has not been assessed in this project.
- *Actual input data (type and format) used*
  The non-TCEQ initial data and the processed intermediate data used to generate the GAMS is contained in the deliverable, as noted in Sections 3.1 and 3.4. The sources of this data are described in Section 2.1.
- *Overview of the immediate (non-manipulated or post-processed) results of the model runs (model application only)*
  The original HYSPLIT back-trajectory model results are contained in the deliverable and described in Section 3.3.2. The post-processed distance and bearing outputs are contained in the intermediate CSV files described in Section 3.3.3 and in the final CSV files described in Section 3.4.2.
- *Output of model runs and interpretation*
  Section 3.6 describes the output files from our GAM fits and cross-validation analysis contained in the deliverable. These results are discussed and interpreted in Sections 2.4, 2.5, and 2.6.
- *User's guide (electronic or paper)*
  This technical memo serves as the user's guide for all the scripts in the deliverable as well as the GAMs provided therein.
- *Instructions for preparing data files (model development only)*
  Input data files for the GAMs must be prepared in a way that matches the format of the final CSV files described in Section 3.4.2 and Appendix A. The units of the variables much match those given in Table 5 (gam01_baseline), Table 8 (gam02_extended), and Table 9 (gam03_extended). The data processing scripts described in Section 3.2 and

contained in the deliverable can be used to prepare these files, but any comma-separated-value file with the necessary columns will work as well.

- *Example problems complete with input and output*
  The input and output of the scripts and GAMs developed in this project are contained in the deliverable and described in Section 3. Section 3.6 describes the output files from our GAM fits and cross-validation analysis contained in the deliverable, which can also be used as example problems.
- *A report of the model calibration, validation, and evaluation (model development only).*
  The calibration of the GAMs, defined as "adjusting model parameters within physically defensible ranges until the resulting predictions give the best possible or desired degree of fit to the observed data," was done as part of the GAM fitting procedure described in Section 2.3. The verification of the GAMs was performed via the two-fold cross-validation described in Section 2.6.

  The evaluation of the HYSPLIT back-trajectories is described in Section 2.2. The GAMs were evaluated as described in Sections 2.4 to 2.6, as well as by addressing the quality assurance questions in Section 4.1.

5 References

Camalier, L., Cox, W., and Dolwick, P. (2007), The effects of meteorology on ozone in urban areas and their use in assessing ozone trends, *Atmos. Environ.*, 41, 7127-7137.

Draxler, R. R. and G. D. Hess (1997), Description of the HYSPLIT_4 modeling system. NOAA Tech. Memo. ERL ARL-224, 24 pp.

Draxler, R. R. and G. D. Hess (1998), An overview of the HYSPLIT_4 modeling system for trajectories, dispersion, and deposition, *Aust. Meteorol. Mag., 47,* 295-308.

Hegarty, H. R. R. Draxler, A. F. Stein, J. Brioude, M. Mountain, J. Eluszkiewicz, T. Nehrkorn, F. Ngan, and A. Andrews (2013), Evaluation of Lagrangian Particle Dispersion Models with Measurements from Controlled Tracer Releases. *J. Appl. Meteor. Climatol.*, *52*, 2623–2637, doi: http://dx.doi.org/10.1175/JAMC-D-13-0125.1

Starkweather, J. (2011), Cross Validation techniques in R: A brief overview of some methods, packages, and functions for assessing prediction models, available at https://www.unt.edu/rss/class/Jon/Benchmarks/CrossValidation1_JDS_May2011.pdf

Wood, S. N. (2006), *Generalized Additive Models: An Introduction with R*, part of the "Texts in Statistical Science" series, Chapman & Hall/CRC, New York.

Appendix A. List of meteorological predictors in *./csv_files/final_files/*

The italicized text below is copied from ./csv_files/final_files/GAMparam_readme.txt in the deliverable.

data is from TCEQ monitoring sites unless indicated:
*** indicates parameter data is from IGRA*
***** indicates parameter data is from NCDC*

Column:
1. Date (YYYYMMDD)
2. MDA8 O3 max (ppbv)
3. MDA8 O3 bkgrd (ppbv)
4. PM2.5 max (ug m-3)
5. PM2.5 bkgrd (ug m-3)

 # TEMPERATURE (C)
6. Maximum surface T
7. Morning average surface T
8. Afternoon average surface T
9. Diurnal T change
10. Minimum Apparent T calculated according to the National Digital Forecast Database algorithm (NDFD)
*11. Maximum Apparent T ****
*12. Average Apparent T ****
*13. 1200 UTC T at 925 mb **
*14. 1200 UTC T at 850 mb **
*15. 1200 UTC T at 700 mb **
*16. 1200 UTC T at 500 mb **
*17. Deviation in T from a 10 year monthly mean at 850 mb **
*18. Deviation in T from a 10 year monthly mean at 700 mb **
*19. Deviation in T from a 10 year monthly mean at 500 mb **
*20. 24h change in 1200 UTC 850 mb T **

 # WIND
21. Average daily u wind vectors
22. Average daily v wind vectors
23. Average daily wind speed (m/s)
24. Average daily wind direction (deg)
25. Morning average u wind vectors
26. Morning average v wind vectors
27. Afternoon average u wind vectors
28. Afternoon average v wind vectors
29. Morning average wind speed (m/s)
30. Morning average wind direction (deg)
31. Afternoon average wind speed (m/s)
32. Afternoon average wind direction (deg)

 # Humidity
*33. Average daily RH (%) ****
*34. Midday average RH (%) ****
*35. Nighttime average RH (%)****
*36. Average dew point T (C) ****
*37. Maximum dew point T (C) ****
*38. Maximum water vapor mixing ratio (g kgô°) ****
*39. Morning 850 mb dew point temperature (C) **
*40. 24-h change in 1200 UTC 850 mb dew point temperatures (C)***

 # Pressure (mb)
*41. Average station pressure (mb) ****
*42. Average sea-level pressure (mb) ****
*43. Morning geo-potential height at 850 mb (m) **
*44. Morning geo-potential height at 700 mb (m) **
*45. Morning geo-potential height at 500 mb (m) **
*46. Deviation in geo-potential height from a 10 year monthly mean at 850 mb (m) **
*47. Deviation in geo-potential height from a 10 year monthly mean at 700 mb (m) **
*48. Deviation in geo-potential height from a 10 year monthly mean at 500 mb (m) **

 # Stability
*49. Difference in 1200 UTC temperatures between surface and 925 (C) **
*50. Difference in 1200 UTC temperatures between surface and 850 (C) **
*51. Maximum afternoon Cloud Ceiling Height (m) ****
 # missing maximum rate of mixing height increase (m h-1)

 # Transport Trajectories
52. 24-h HYSPLIT transport direction (degrees from N)
53. 24-h HYSPLIT transport distance (km)

 # Synoptic Weather
54. Solar Radiation (Langy/min) # replacing average morning and afternoon fractional cloud cover (%)
*55. Total precipitation (in) ****
*56-59. Binary indicators of the occurrence of rain, haze, and fog - see readme.txt in /TCEQ/NCDC_ISH/ directory for definitions ****