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

FINAL REPORT

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

TCEQ Contract No. 582-15-50415
Work Order No. 582-15-54118-01
Revision 2.0

Prepared by:
Matthew Alvarado, Chantelle Lonsdale, Marikate Mountain, and Jennifer Hegarty
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

August 31, 2015

Document Change Record

Revision	Revision Date	Remarks
1.0	**15 August 2015**	**Draft version submitted to TCEQ**
2.0	**31 August 2015**	**Final version submitted to TCEQ**

TABLE OF CONTENTS

List of Figures

List of Tables

List of Acronyms

AER – Atmospheric and Environmental Research
AERONET – Aerosol Robotic NETwork
ANOVA – ANalysis Of VAriance
AOD – Aerosol Optical Depth
ARR – Austin/Round Rock
BG – Background
BPA – Beaumont/Port Arthur
CAMx – Comprehensive Air Quality Model with Extensions
CMAQ – Community Multi-scale Air Quality Model
CSV – Comma Separated Value
CTM – Chemical Transport Model
DFW – Dallas/Fort Worth
GAM – Generalized Additive Model
GEO-CAPE – GEOstationary Coastal and Air Pollution Events
GLM – Generalized Linear Model
HYSPLIT – Hybrid Single Particle Lagrangian Integrated Trajectory Model
IQR – Inter-Quartile Range
LIDORT – Linearized Discrete ORdinate Radiative Transfer model
MDA8 – maximum daily 8-hour average ozone
MODIS – Moderate Resolution Imaging Spectroradiometer
MT – Map Type
NAAQS – National Ambient Air Quality Standards
NCEP – National Centers for Environmental Prediction
O_3 – Ozone
OE – Optimal Estimation
OLS – Ordinary Least Squares
OMI – Ozone Monitoring Instrument
OSSE – Observing System Simulation Experiment
PC1 – Principle Component 1
PCA – Principle Component Analysis
$PM_{2.5}$ – Particulate Matter with diameter below 2.5 microns
ppbv – Parts Per Billion by Volume
QAPP – Quality Assurance Project Plan
RH – Relative Humidity
SA – San Antonio

SO_2 – Sulfur Dioxide
TCEQ – Texas Commission on Environmental Quality
TES – Tropospheric Emission Spectrometer
TLM – Tyler/Longview/Marshall
URBE – Un-Biased Risk Estimator
UTC – Coordinated Universal Time

Executive Summary

In this project, AER investigated the impact of meteorology on O_3 and $PM_{2.5}$ in six urban areas in Texas (Houston/Galveston/Brazoria, Dallas/Fort Worth, San Antonio, Austin/Round Rock, Beaumont/Port Arthur, and Tyler/Longview/Marshall). The purpose of this project was to:

1) Investigate the temporal trends of regional background O_3 and $PM_{2.5}$ and

2) Determine what synoptic- and urban-scale meteorological conditions are important in explaining and forecasting high concentrations of O_3 and $PM_{2.5}$.

To accomplish this, we first estimated daily regional background concentrations of O_3 and $PM_{2.5}$ for a ten-year period (2005-2014) for six urban areas and for the State of Texas as a whole using both the TCEQ method (lowest value at a set of background sites) and a principal component analysis (PCA) based technique. We then derived updated generalized additive models (GAMs) relating urban total and background O_3 and $PM_{2.5}$ to urban-scale meteorological predictors. We find that urban-scale meteorological predictors can explain 65-80% of the variability in urban O_3, but only 30-40% of the variability in urban $PM_{2.5}$. After using these relationships to correct the observed trends in total and background O_3 and $PM_{2.5}$, we find significant (95% confidence) negative trends in all four pollution metrics for Houston/Galveston/Brazoria and Dallas/Fort Worth, as well as significant negative trends in some of the pollution metrics for the other urban areas. However, the meteorologically adjusted trends in the background O_3 and $PM_{2.5}$ are similar to the meteorologically adjusted trends in the total, suggesting most of the observed trend in total O_3 and $PM_{2.5}$ is due to trends in the regional background rather than changes in local production. The major exception is O_3 in SA, which shows a trend near zero in total O_3 but a significant negative trend of about -1.0 ppbv/year in background O_3, suggesting that local O_3 production may have increased in SA between 2005-2014.

Our seasonal trends analysis shows that background O_3 is fairly constant through the O_3 season (May-October) in Dallas/Fort Worth and Tyler/Longview/Marshall, but has a minimum in July for the other urban areas. In contrast, background $PM_{2.5}$ peaks in June and July in all urban areas.

We then determined a set of five synoptic "types" covering 70% of all days (and 58% of the days in the O_3 season) for our ten-year study period. We found that the relative frequency of high O_3 and $PM_{2.5}$ events in the urban areas varied significantly with these synoptic types. We used logistic regression to develop a model that predicted the probability of high O_3 and $PM_{2.5}$ events as a function of three variables: the derived synoptic map types, afternoon mean temperature, and daily mean wind speed. We used these functions to determine necessary and sufficient conditions for high O_3 and $PM_{2.5}$ events, and determined that these conditions vary significantly between the urban areas considered here.

We recommend that future work focus on:

1) Developing additional methods to determine regional background concentrations, including further research into the differences between the TCEQ and PCA-based background estimates for O_3 and the use of satellite observations, combined with chemical transport models, to determine background $PM_{2.5}$,
2) Further investigations into the synoptic types controlling high O_3 and $PM_{2.5}$ in Texas, and

3) Further developing the GAMs and logistic models developed in this project to forecast air quality in Texas urban areas and evaluate the ability of 3D Eulerian air quality models to correctly simulate the impact of meteorology on O_3 and $PM_{2.5}$.

1. Introduction

1.1 Project Objectives

AER performed a research project titled "Investigating the Impact of Meteorology on O_3 and $PM_{2.5}$ Trends, Background Levels, and NAAQS Exceedances" for the Texas Commission on Environmental Quality (TCEQ). The objectives of this project were to:

- Determine the effects of meteorology on trends in O_3 and $PM_{2.5}$ by developing new generalized additive models (GAMs) for O_3 and $PM_{2.5}$ concentrations to selected meteorological variables for the urban areas in Table 1.
- Estimate the regional background concentrations of O_3 and $PM_{2.5}$ for the urban areas in Table 1.
- Investigate the synoptic and urban-scale meteorological conditions that are associated with (i.e., are necessary and/or sufficient for) high concentrations of background and total O_3 and $PM_{2.5}$ in the "Group 1" urban areas in Table 1.

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 schedule of deliverables for this project is given in Table 2, while the purpose and background of each of the three tasks is summarized below.

1.2 Purpose and Background

1.2.1 Trends in O_3 and $PM_{2.5}$

As the formation and loss of pollutants such as O_3 and $PM_{2.5}$ are strongly influenced by meteorology, inter-annual trends in these pollutants represent a combination of changes due to inter-annual variability in meteorology and changes due to air quality policy actions and other economic and societal trends. Statistical techniques are thus used to account for the effect that meteorological variations have on the trends of O_3 and $PM_{2.5}$ so that the adjusted trends can be used to assess the effectiveness of air quality policy. A common approach to performing this "meteorological adjustment" is to use a generalized additive model (GAM, Wood, 2006) to describe the potentially non-linear relationship between measured O_3 (maximum daily 8-hour average, or MDA8) or $PM_{2.5}$ (daily average) concentrations and selected meteorological variables (e.g., Camalier et al., 2007). In this project, AER derived updated GAMs for urban O_3 and $PM_{2.5}$ for the urban areas in Table 1. AER used these models to account for the effect that meteorological variations have on the trends of O_3 and $PM_{2.5}$.

1.2.2 Regional Background Concentrations of O_3 and $PM_{2.5}$

Daily surface concentrations of O_3 and $PM_{2.5}$ in urban areas can be considered as the sum of O_3 and $PM_{2.5}$ produced within the urban area (either through primary emissions of $PM_{2.5}$ or through secondary chemical production of O_3 and $PM_{2.5}$) and a "regional background" that is

transported into the urban area. Accurate estimates of this regional background are critical to determining the potential for further reductions in O_3 and $PM_{2.5}$ concentrations in urban areas through control of local emissions of primary $PM_{2.5}$ and the precursors of O_3 and $PM_{2.5}$.

In this project, AER determined daily regional background estimates of O_3 and $PM_{2.5}$ for a ten-year period (2005-2014) for the urban areas in Table 3 and for the State of Texas as a whole using the TCEQ method (i.e., the lowest value observed at defined "background" sites near the border of the area of interest, Berlin et al., 2013). AER then used these background estimates to investigate the spatial and temporal trends of regional background O_3 and $PM_{2.5}$.

AER also explored other data-based ways of determining regional background concentrations of O_3 and $PM_{2.5}$ (e.g., the principle component analysis method of Langford et al., 2012, or the use of satellite observations of O_3 and aerosols) using data from the "Group 1" urban areas.

1.2.3 Synoptic- and Urban-scale Meteorological Controls on O_3 and $PM_{2.5}$

In this project AER investigated what synoptic- and urban-scale meteorological conditions are important in explaining and forecasting high concentrations of O_3 and $PM_{2.5}$ in the "Group 1" urban areas listed in Table 1. We identified necessary and/or sufficient meteorological conditions that lead to high concentration events (e.g., above 90^{th} percentile) for these pollutants. Meteorological conditions leading to both high regional background levels and high total levels of O_3 and $PM_{2.5}$ were identified.

1.3 Report Outline

This Final Report documents the methods and pertinent accomplishments of this project, including comprehensive overviews of each task, a summary of the data collected and analyzed during this work, key findings, shortfalls, limitations and recommended future tasks. It satisfies Deliverable 5.2 of the Work Plan for Work Order No. 582-15-54118-01

> **Deliverable 5.2:** Final Report delivered electronically via file transfer protocol or e-mail in Microsoft Word format and PDF format
>
> **Deliverable Due Date:** August 31, 2015

This report contains three sections that describe the methods and major findings for Task 2 (Effects of Meteorology on O_3 and $PM_{2.5}$ Trends, Section 2), Task 3 (Estimating Background O_3 and $PM_{2.5}$, Section 3) and Task 4 (Importance of Synoptic/Mesoscale Meteorological Conditions in Explaining/Forecasting Maximum O_3 and $PM_{2.5}$, Section 4). Technical memos previously delivered to TCEQ relating to Tasks 2 and 3 are included as Appendices A and B, respectively, while Appendix D includes additional plots relating to Task 4.

Section 5 discusses the Quality Assurance performed for the project, including answers to the assessment questions from the Quality Assurance Project Plan (QAPP). Section 6 summarizes the conclusions of our study, and Section 7 lists our recommendations for further research.

In addition, Appendix C describes the files that are included in the final deliverable package (Deliverable 5.2).

Table 2. Projected Schedule for TCEQ Work Order No. 582-15-54118-01

Milestones	Planned Date
Task 1 - Work Plan	
1.1: TCEQ-approved Work Plan	April 3, 2015
1.2: TCEQ-approved QAPP	April 3, 2015
Task 2 - Effects of Meteorology on O_3 and $PM_{2.5}$ Trends	
2.1: Monthly teleconferences or meetings to deliver project status to the TCEQ Project Manager	Monthly
2.2: Deliver a technical memo describing GLMs relating meteorological variables to measured MDA8 O3 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 shall also attach R scripts and other computer codes used to generate and/or analyze the GLMs.	June 30, 2015
Task 3 – Estimating Background O_3 and $PM_{2.5}$	
3.1: Daily estimates of regional background O_3 (May through October) and $PM_{2.5}$ (all year) by the TCEQ's method for 2005-2014 for the Group 1 and Group 2 metropolitan areas, as well as the state of Texas.	May 29, 2015
3.2: Deliver, as part of the draft and final reports, an analysis of the spatial and temporal trends in the estimates of regional background O_3 and $PM_{2.5}$.	Draft: August 15, 2015 Final: August 31, 2015
3.3: Deliver, as part of the draft and final reports, an analysis of alternative data-based methods to determine regional background O_3 and $PM_{2.5}$.	Draft: August 15, 2015 Final: August 31, 2015
Task 4 – Importance of Synoptic/Mesoscale Meteorological Conditions in Explaining/Forecasting Maximum O_3 and $PM_{2.5}$	
4.1: Deliver, as part of the draft and final reports, a description of synoptic map "types" associated with high levels of background and total O_3 and $PM_{2.5}$ for the Group 1 metropolitan areas.	Draft: August 15, 2015 Final: August 31, 2015
4.2: Deliver, as part of the draft and final reports, a description of urban-scale meteorological predictors of O_3 and $PM_{2.5}$ exceedances for the Group 1 metropolitan areas.	Draft: August 15, 2015 Final: August 31, 2015
Task 5 – Draft and Final Reports	
5.1: Draft Report for TCEQ review and approval, delivered electronically via file transfer protocol or e-mail in Microsoft Word format and PDF format	August 15, 2015
5.2: Final Report delivered electronically via file transfer protocol or e-mail in Microsoft Word format and PDF format	August 31, 2015

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

The technical memo delivered on June 30, 2015 described the generalized additive models (GAMs) that related meteorological variables to measured MDA8 O_3 and $PM_{2.5}$ for urban areas in Table 1 on data for the O_3 season (May through October) from 2005-2014 and $PM_{2.5}$ from 2005-2014. It is discussed and summarized below.

2.1 Input Data and Processing

The procedure in which we derived the daily minimum and maximum MDA8 ozone, and $PM_{2.5}$, meteorological parameters is described in detail in Appendix A. Effects of Meteorology on O_3 and $PM_{2.5}$ Trends and Appendix B. Technical Memo: Estimating Background O_3 and $PM_{2.5}$. Figure A.16 and Figure A.17 displays the order of functions and scripts that take the ozone, $PM_{2.5}$ and meteorological measurement data (Section A.2) to the CSV-ready files used in the GAM modeling function and background analysis. The raw input data sets are described in Section A.1.1, the HYSPLIT Back Trajectories for each urban region is described in Section A.2.3, the calculation of MDA8 ozone and hourly $PM_{2.5}$ is described in Section AB.2.2, and estimates of the TCEQ method background is described in Section AB.2.3.

2.2 Generalized Additive Model

As the formation and loss of O_3 and $PM_{2.5}$ are strongly influenced by meteorology, inter-annual trends in these pollutants represent a combination of changes due to inter-annual variability in meteorology and changes due to air quality policy actions and other economic and societal trends. Statistical techniques are thus used to account for the effect that meteorological variations have on the trends of O_3 and $PM_{2.5}$ so that the adjusted trends can be used to assess the effectiveness of air quality policies. A common approach to performing this "meteorological adjustment" is to use a generalized additive model (GAM, Wood, 2006) to describe the potentially non-linear relationship between measured urban O_3 MDA8 or $PM_{2.5}$ (daily average) concentrations and selected meteorological variables taken from an array of candidate meteorological variables (e.g., Camalier et al., 2007). TCEQ has developed such models in the past, but these models have not been updated since 2008.

The easiest way to understand the GAM approach is to contrast it with two related, but simpler, approaches: ordinary linear models and generalized linear models. In an ordinary linear model (e.g., Wood, 2006, p. 12), the model equation is:

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} \quad \mathbf{y} \sim N(\boldsymbol{\mu}, \boldsymbol{I_n}\sigma^2)$$

where $\boldsymbol{\mu}$ is a vector of the expected values of the observation vector, $\mathbf{y}$, (both of dimension N_{obs}), which is assumed to be normally distributed around the expected values with a constant variance of σ^2. $\mathbf{X}$ is a matrix of predictor variables (dimension N_{obs} by N_{preds}), and $\boldsymbol{\beta}$ is the (initially unknown) vector of best-fit coefficients for the predictor variables. Note that this functional form is not as limited as it first appears. For example, known non-linear functions of the predictor variables (e.g., $x_i^2, \sin\frac{x_i^2}{x_j^3}$) can be used as new predictor variables, and the observation vector $\mathbf{y}$ can be similarly transformed to make it normally distributed (e.g., taking the logarithm of a log-normally distributed observation).

However, ordinary linear models have two inherent limitations. The first is the requirement that the observation be distributed according to a normal distribution. This rules out the use of ordinary linear models to predict observations that follow other distributions, such as when you

wish to predict the probability that the result of an experiment will be true or false based on a set of predictors (e.g., logistic regression), and thus your observations are expected to follow a binomial distribution. Generalized linear models (Wood, 2006, p. 59) relax this normality requirement so that distributions of any exponential family (Poison, Binomial, Gamma, Normal) can be used, as well as a set of "link" functions – smooth, monotonic functions of the expected value vector **μ**.

The second limitation of ordinary (and generalized) linear models is that they require that the functional dependence of the observation on the predictor variables be specified ahead of time, with only the linear coefficients **β** of those functions allowed to vary. This makes these approaches less useful where the functional form of the response is not known, or where it might be highly complex. In this case, a generalized *additive* model can be used (Wood, 2006, p. 121). The response of each predictor variable is expected to be a non-linear but smooth function constructed as a linear sum of group of simpler basis functions of the predictor. By fitting the coefficients of these basis functions, one can estimate the previously unknown smooth function of the predictor. Cubic splines are generally used as the basis functions, as this ensures the resulting smooth function is continuous up to the second derivative.

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 (Wood, 2006) in R (R Core Team, 2015). 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 *i*th 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.

2.2.1 Baseline GAMs (gam01_baseline)

We 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 A.4 with the results discussed in Section A.1.4.

2.2.1.1 MDA8 O_3 Results Summary

- 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 greater than 30 °C, 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 approximately 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 back-trajectory bearing of approximately 150° (southeast), likely due to reduced background O_3 from flows from the Gulf of Mexico.
- The day-of-year function shows a minimum at approximately 200 Julian days (July) in each urban area.

2.2.1.2 Maximum Daily Average $PM_{2.5}$ Results Summary

- 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 greater than approximately 30 °C.
- The fits for mid-day RH are very uncertain at low (less than 40%) and high (greater than 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 approximately 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 approximately 60° (northeast) and a minimum around 320° (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 approximately 150° (southeast), likely due to flows from the Gulf of Mexico transporting dust from North Africa into the urban areas.
- 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 approximately 200 Julian days (July), which again may be related to the transport of North African dust into the urban area from the Gulf of Mexico.

2.2.2 Extended GAMs (gam02_extended and gam03_extended)

We also explored whether a different set of meteorological predictors than those used by Camalier et al. (2007) and used in the baseline GAMs of A.1.4 could provide a better fit to the maximum MDA8 O_3 and maximum daily average $PM_{2.5}$ for each urban area. The procedure we used is described in detail in Section A.1.5.

2.2.2.1 MDA8 O_3 Results Summary

- 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 greater than 30 °C, 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 °C, 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*); at BPA, the effect of this predictor was found to be insignificant and so was dropped from the final model. However, O_3 decreases at the highest values of *T_diff_850mb* for SA (-5 to 0 °C).
- 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 approximately 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 approximately 150° (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 approximately 200-225 Julian days (July-August).

2.2.2.2 Maximum Daily Average $PM_{2.5}$ Results Summary

- 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 greater than approximately 30 °C.
- 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 (approximately 20%).
- $PM_{2.5}$ generally increases with increasing temperature at 925 mbar, but HGB also shows a possible increase in $PM_{2.5}$ at low 925 mbar temperatures.
- $PM_{2.5}$ 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 approximately 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 back-trajectory bearing of approximately 60° (northeast). DFW, SA, ARR, and TLM show a minimum around 320° (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 approximately 150° (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 approximately 200 Julian days (July).

2.2.3 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 that 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}).

Full details of this cross-validation are described in Section A.1.6.

Table 3 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 (less than 1 ppbv for O_3 and less than 0.25 μg 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 over-fitting 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. A list of suspicious predictors based on this analysis is included in Table A.13.

Table 3. 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

2.3 GAMs for Background $PM_{2.5}$ and O_3

We also used the same approach used to derive the *gam03* models described in Section 2.2.2 to fit GAMs for the background MDA8 O_3 and daily average $PM_{2.5}$ for the four Group 1 urban areas. These models, called *back_gam03* here for convenience, will be provided to TCEQ as part of the final deliverable package described in Appendix C and are used to calculate meteorologically adjusted trends in background O_3 and $PM_{2.5}$ in Section 2.4. One thing to note is that, while the model intercept (β_o) and year-to-year variability terms (Y_k) differ between the models fit to total and background pollutant values, the shape and magnitude of the smooth functions for the meteorological predictors is remarkably consistent between the *gam03* and *back_gam03* models. The only noticeable difference is for O_3 in the HGB area, where the GAM for total O_3 shows a stronger dependence on mean afternoon temperature and daily average wind speed than the background O_3 GAM, suggesting that those predictors have a strong influence on local chemical production of O_3 in HGB. Further work should attempt to fit measures of locally produced O_3 and $PM_{2.5}$ (i.e., total minus background) to meteorological predictors to see if these dependences differ significantly from those for regional background and total O_3 and $PM_{2.5}$.

2.4 Meteorologically Adjusted Trends of O_3 and $PM_{2.5}$

We used the "extended" *gam03* models described in Sections 2.2.2 and 2.3 to determine the meteorologically adjusted trends in total and background MDA8 O_3 and daily average $PM_{2.5}$. In this procedure, we use the Y_k terms from the GAM equation in Section 2.2 to determine the

relative difference between the annual averages after meteorology has been taken into account. Our equation for the annual averages is thus

$$g(\mu_k) = \beta_o + Y_k + c_o$$

where k is the k^{th} year's average and c_o is a constant. The constant c_o is needed because of how R treats factor variables. In order to have an identifiable model, one of the factor levels, in this case the year 2005, must be set to have a value of $Y_k = 0$. However, the year 2005 is frequently the year with the largest annual average O_3 and $PM_{2.5}$ values in the original data set. This results in Y_k values that are predominantly less than 0, leading to meteorologically adjusted annual averages that do not have the same 10-year average as the original data set. To avoid this issue, we add a constant c_o to the meteorologically adjusted annual averages so that the 10-year averages in the original and meteorologically adjusted trend data are identical. The value of the meteorologically adjusted linear trends over 2005-2015 is relatively insensitive to the value of c_o.

The original and meteorologically adjusted annual averages are shown in Figure 1 through Figure 6 below. The trend estimates, determined by ordinary least squares (OLS) linear regression of the annual averages, are summarized in Table 4 below. As expected, the meteorologically adjusted annual averages show less year-to-year variation and generally show trends closer to zero than the original data. The largest impact of the meteorological adjustment is on the trends for O_3 in HGB and BPA. In general, the meteorological adjustment affects estimates of O_3 trends more than $PM_{2.5}$ trends – this is to be expected as the GAMs account for a larger fraction of the observed variability in O_3 than for $PM_{2.5}$.

No positive trends with time are observed for any of the six urban areas examined here for 2005-2014 either before or after meteorological adjustment. The meteorologically adjusted negative trends are significant at an $\alpha = 0.05$ level for all pollutant metrics at HGB and DFW. The adjusted trends are also significant for total MDA8 O_3 at TLM, for total $PM_{2.5}$ at BPA and TLM, and for background $PM_{2.5}$ at SA and ARR. However, the meteorologically adjusted trends in the background O_3 and $PM_{2.5}$ for the four Group 1 urban areas are similar to the meteorologically adjusted trends in the total, suggesting most of the observed trend in total O_3 and $PM_{2.5}$ is due to trends in the regional background rather than changes in local production. The major exception is O_3 in SA, which shows a trend near zero in total O_3 but a significant negative trend of -1.01±0.87 ppbv/year in background O_3, suggesting that local O_3 production may have increased in SA between 2005-2014. HGB and DFW show slightly larger decreases in total O_3 than in background O_3, suggesting that local production has decreased in these urban areas.

Table 4. Original and meteorologically adjusted linear trends (±95% confidence intervals) of total and background (BG) MDA O_3 and daily average $PM_{2.5}$ from 2005-2015 using the *gam03* models. Trends significantly different from zero with 95% confidence are in bold. NA is used for Group 2 urban areas where background GAMs were not fit, and so meteorologically adjusted trends were not calculated.

Urban Area	Total MDA8 O_3 (ppbv/year)		BG MDA8 O_3 (ppbv/year)		Total Daily $PM_{2.5}$ ($\mu g/m^3$/year)		BG Daily $PM_{2.5}$ ($\mu g/m^3$/year)	
	Orig.	Met. Adj.	Orig.	Met. Adj.	Orig.	Met. Adj.	Orig.	Met. Adj.
HGB	**-1.48±0.73**	**-0.46±0.40**	**-0.91±0.57**	**-0.41±0.31**	**-0.55±0.13**	**-0.39±0.14**	**-0.48±0.11**	**-0.37±0.11**
DFW	-0.81±0.90	**-0.68±0.39**	-0.61±0.84	**-0.57±0.49**	**-0.17±0.10**	**-0.15±0.09**	**-0.24±0.10**	**-0.23±0.08**
SA	-0.02±0.85	0.00±0.56	-1.00±1.04	**-1.01±0.87**	-0.08±0.09	-0.08±0.14	**-0.10±0.08**	**-0.12±0.07**
ARR	-0.44±0.75	-0.35±0.42	-0.33±0.98	-0.36±0.56	**-0.10±0.09**	-0.10±0.12	**-0.21±0.08**	**-0.21±0.10**
BPA	**-1.03±0.65**	-0.15±0.55	**-1.13±0.62**	NA	**-0.48±0.11**	**-0.34±0.11**	**-0.29±0.13**	NA
TLM	-0.78±1.05	**-0.66±0.41**	-0.65±1.10	NA	**-0.34±0.15**	**-0.34±0.08**	**-0.34±0.15**[a]	NA

[a]Note that at TLM, the total and background $PM_{2.5}$ estimates are identical (see Table B.3).

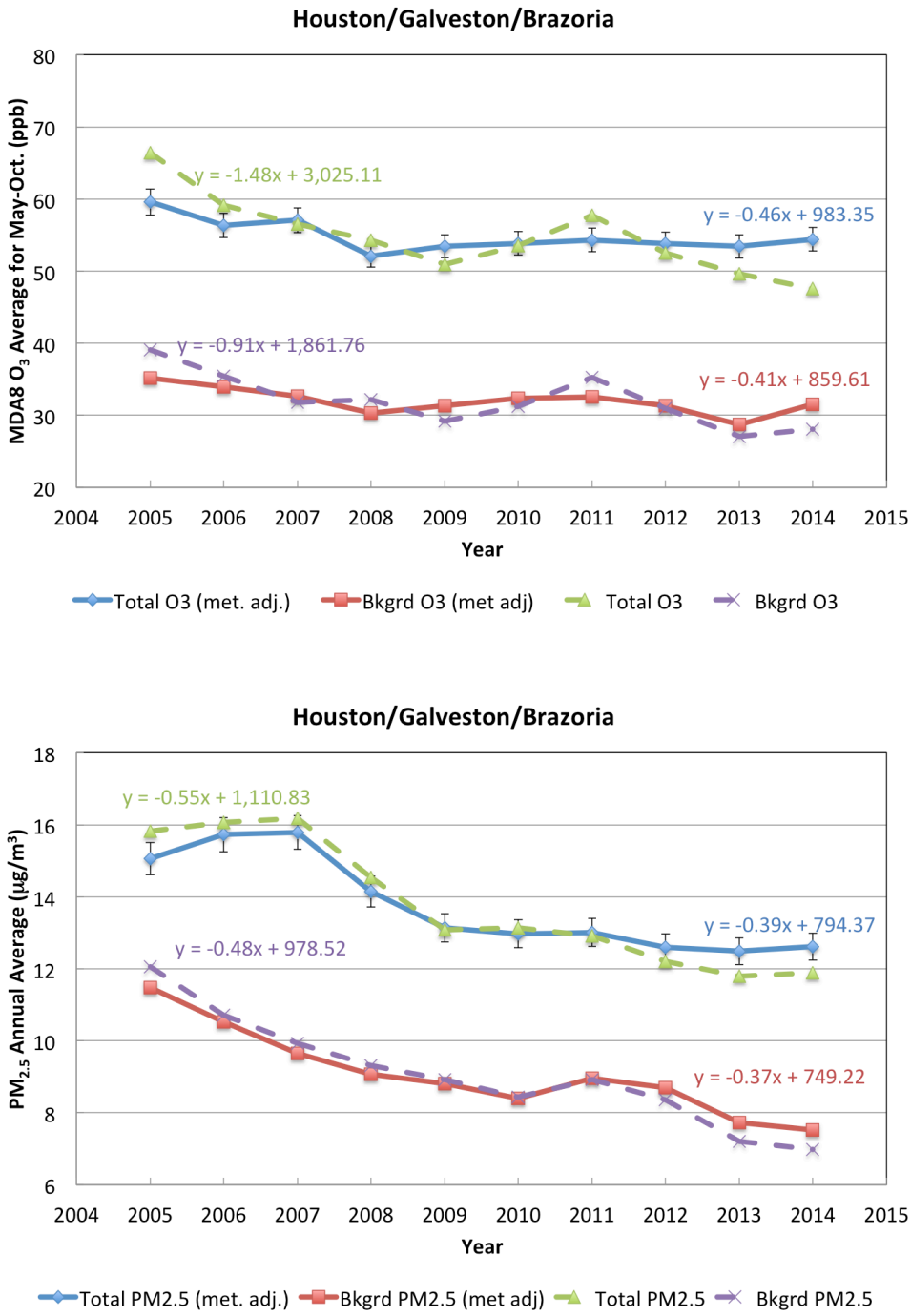


Figure 1. Original (dashed lines) and meteorologically adjusted (solid lines) annual averages for total and background O_3 (top) and $PM_{2.5}$ (bottom) for the Houston/Galveston/Brazoria urban area. Equations for the OLS linear regressions are shown on the plot as well.

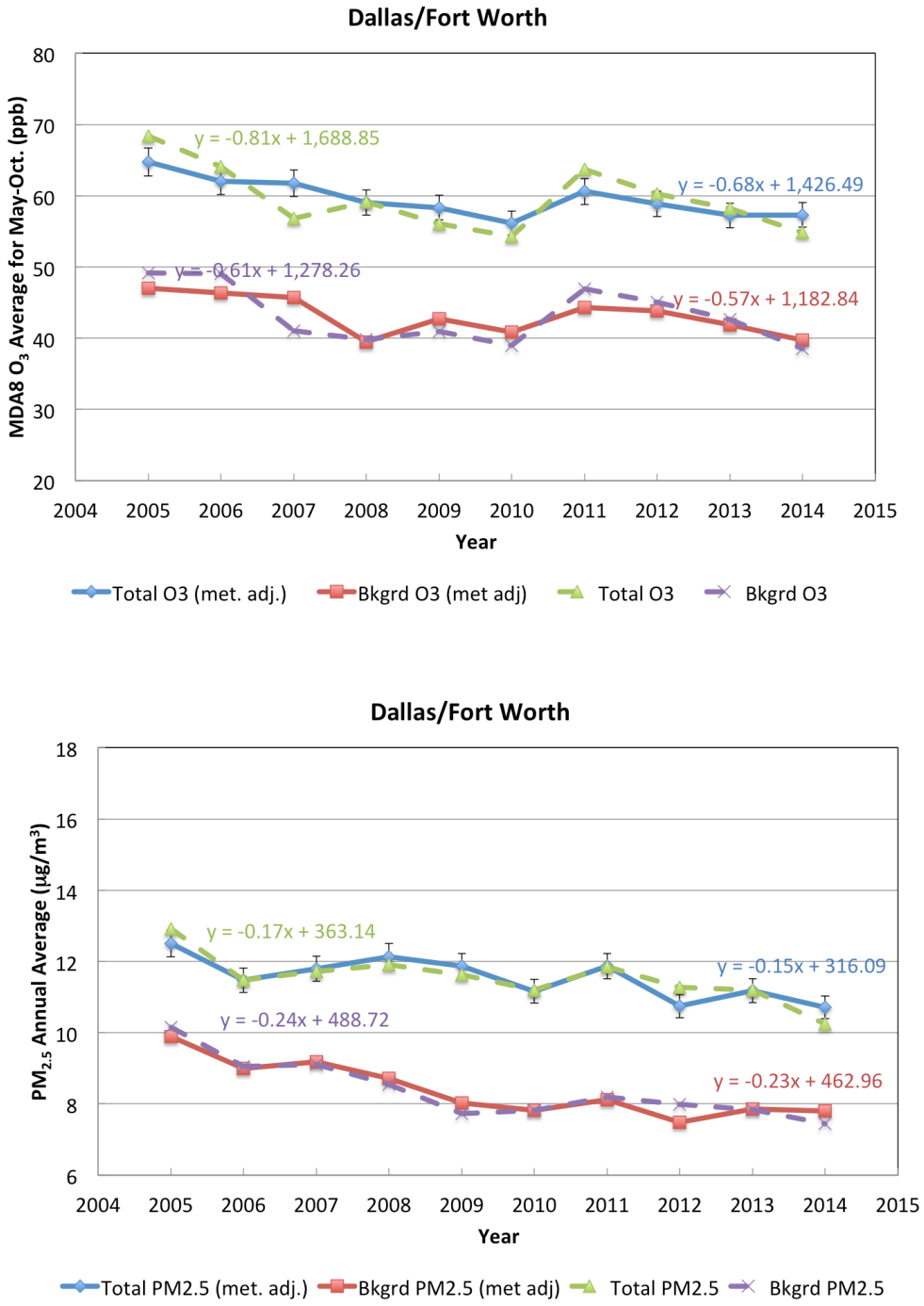
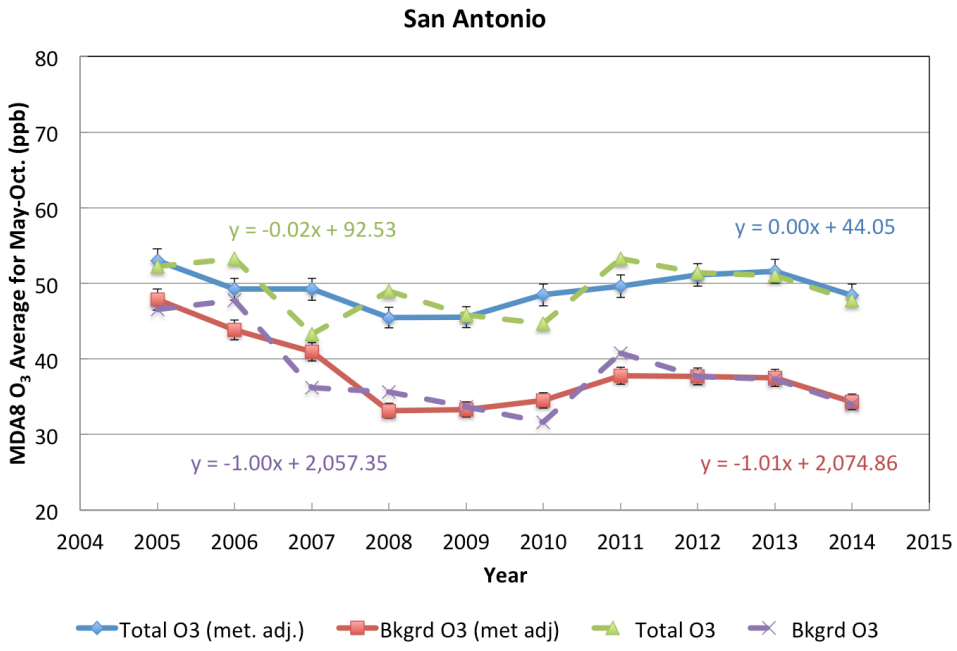


Figure 2. As in Figure 1 but for the Dallas/Fort Worth urban area.



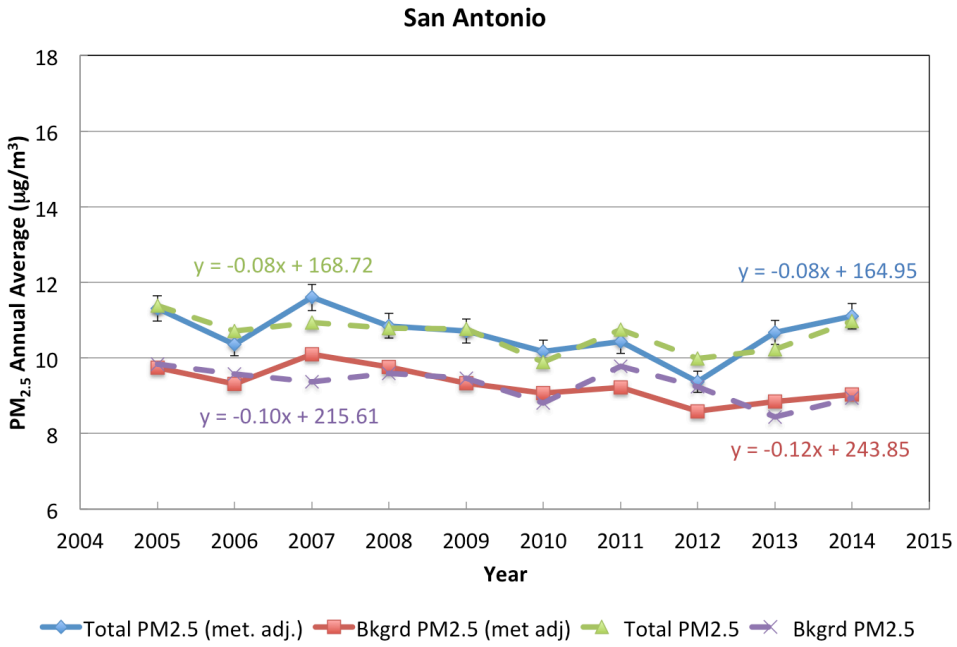


Figure 3. As in Figure 1 but for the San Antonio urban area.

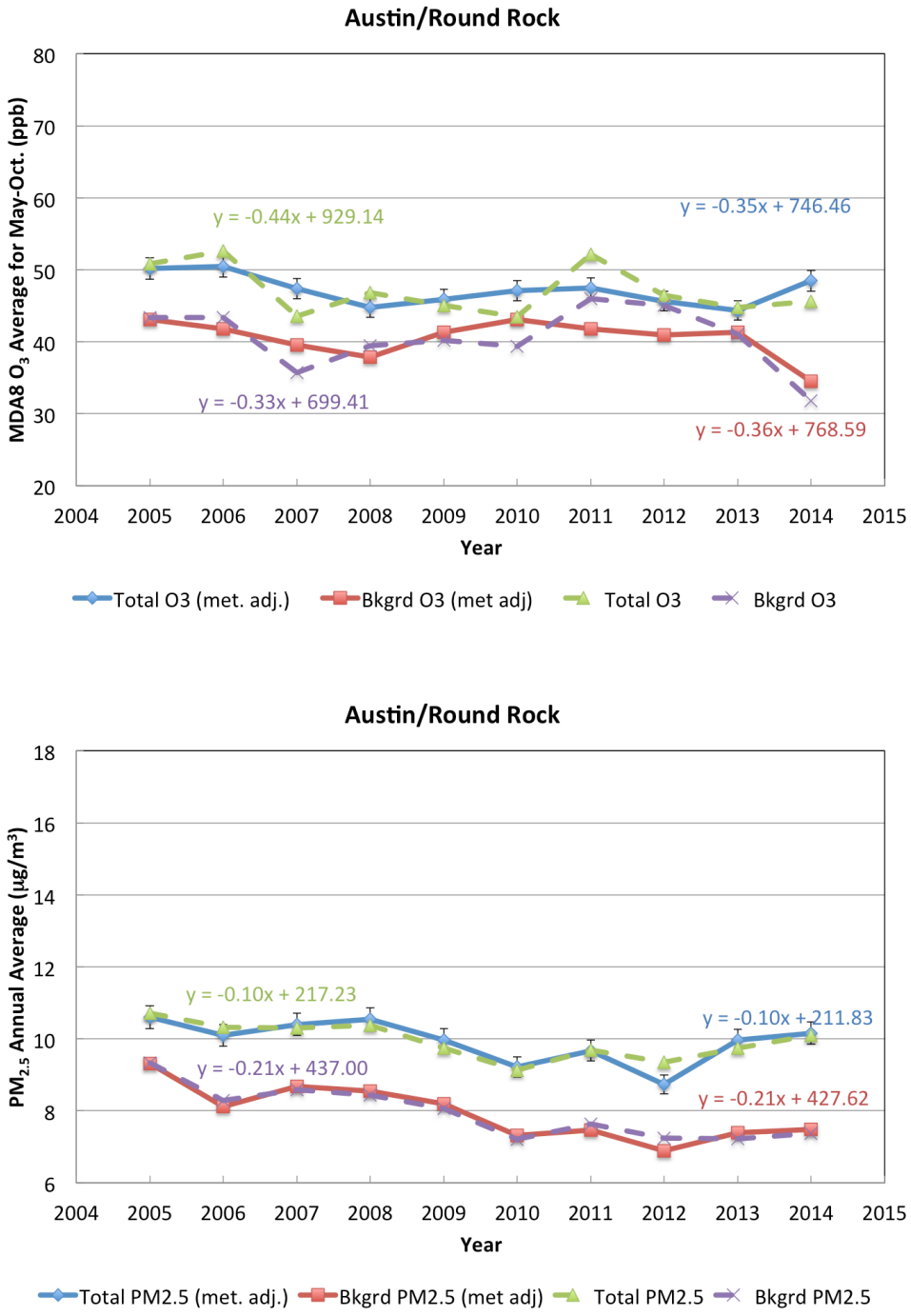


Figure 4. As in Figure 1 but for the Austin/Round Rock urban area.

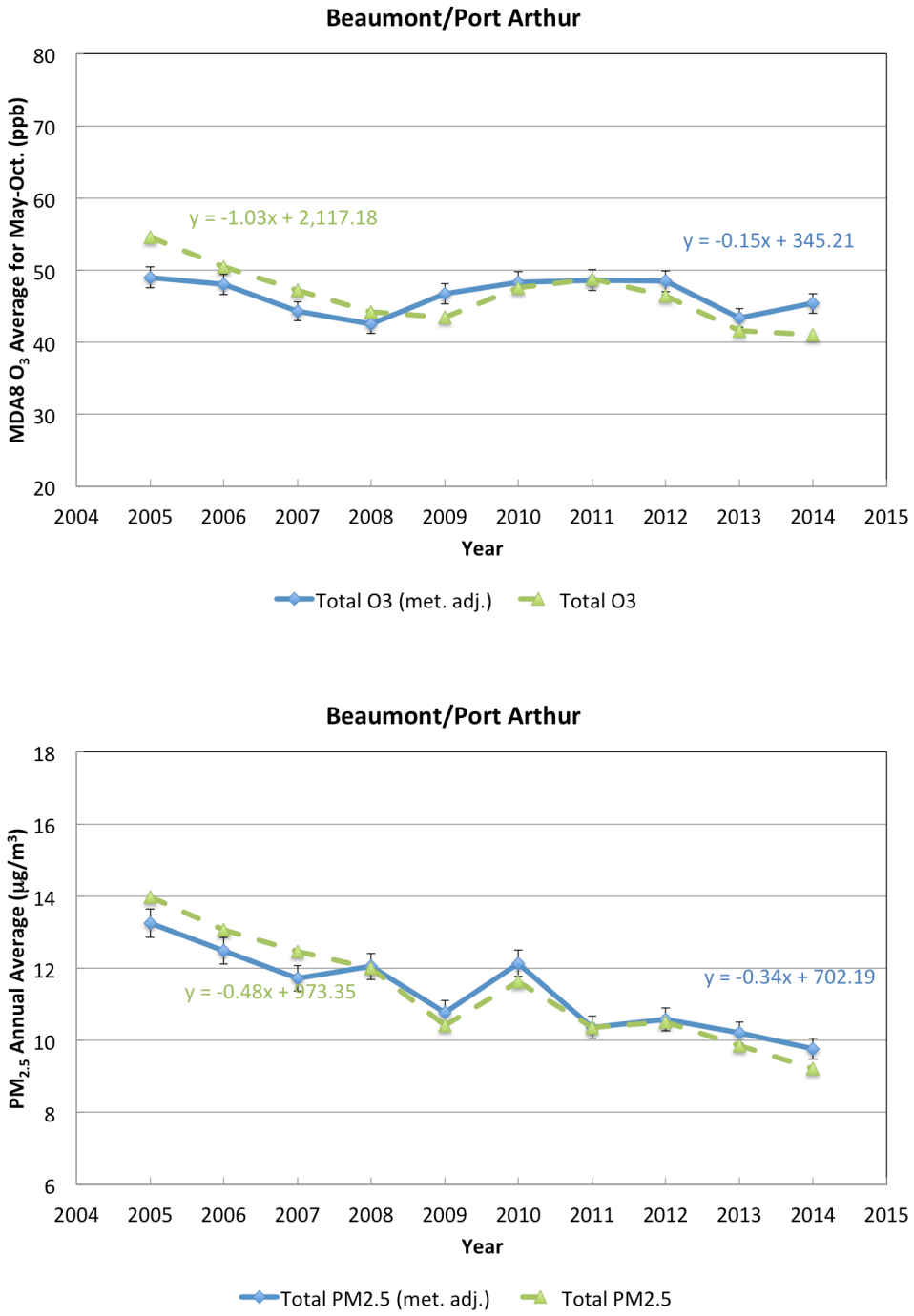


Figure 5. Original (dashed lines) and meteorologically adjusted (solid lines) annual averages for total O_3 (top) and $PM_{2.5}$ (bottom) for the Beaumont/Port Arthur urban area. Equations for the OLS linear regressions are shown on the plot as well.

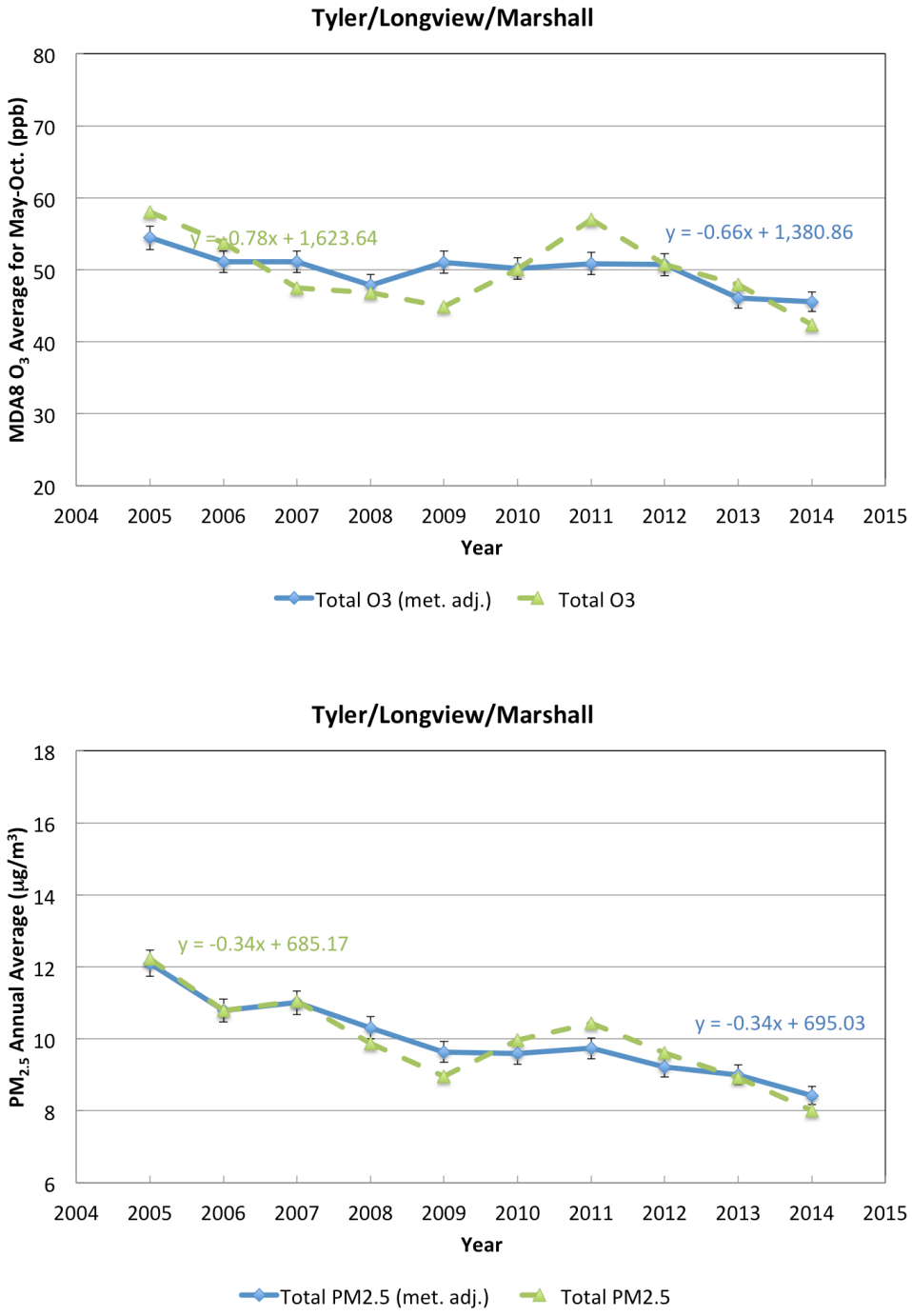


Figure 6. As in Figure 5 but for the Tyler/Longview/Marshall urban area.

2.5 Conclusions

- The Generalized Additive Model (GAMs) relating meteorological variables to the maximum MDA8 O_3 for each urban area generally explain 65-80% of the deviance[1] (i.e. variability), consistent with the results of Camalier et al. (2007). The 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 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 over-fitting.
- However, the cross validation analysis also shows that the smooth function fit for each meteorological predictor 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.
- 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 °C 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° instead of at 320°.
- Similarly, we find that the meteorological relationships fit to background O_3 and $PM_{2.5}$ are substantially identical to those fit to total O_3 and $PM_{2.5}$, with the possible exception of HGB O_3. In HGB, the GAM for total O_3 shows a stronger dependence on mean afternoon temperature and daily average wind speed than the background O_3 GAM, suggesting that those predictors have a strong influence on local chemical production of O_3 in HGB.
- No positive trends with time are observed for any of the six urban areas examined here for 2005-2014 either before or after meteorological adjustment. The meteorologically adjusted negative trends are significant at an $\alpha = 0.05$ level for all pollutant metrics at HGB and DFW. The adjusted trends are also significant for total MDA8 O_3 at TLM, for background O_3 at SA, for total $PM_{2.5}$ at BPA and TLM, and for background $PM_{2.5}$ at SA and ARR. The results suggest most of the observed trends in total O_3 and $PM_{2.5}$ are due to trends in the background rather than changes in local production, with the exception of O_3 in SA (where local production appears to be increasing) and HGB and DFW (where local production appears to be decreasing).

[1] "Deviance" plays a similar role as the variance of the residuals in linear models (Wood, 2006, p. 70). The percent of deviance explained is a generalization of r^2 from linear models.

3 Task 3: Background O_3 and $PM_{2.5}$

3.1 Daily Estimates of Regional Background O_3 and $PM_{2.5}$ (TCEQ Method)

The detailed description of our application of the TCEQ method to derive background O_3 and $PM_{2.5}$ concentrations are discussed in Appendix B. This method selects the lowest valid measured value at a set of "background" sites around the urban area as the background estimate.

We performed a linear regression quality check of these results as discussed in detail in Section AB.2. Figure B.1 shows a scatterplot of the background MDA8 O_3 value versus the maximum MDA8 O_3 value for the HGB area. The solid black line is the linear fit, and the dotted and dashed black lines are the upper and lower 95% (or 2σ) confidence intervals, respectively. In this example, 89 of the 1834 valid data points (4.9%) have maximum MDA8 O_3 values that fall above the upper confidence internal of the linear fit, suggesting that these background estimates are lower than would be expected given the maximum values seen in the urban area. Table B.3 gives the number of such points for each urban area and pollutant.

Similar to Berlin et al. (2013), we performed further analysis of the points that were above the 95% confidence interval of the fit (e.g., where high_flag = TRUE). First, we identified the subset of those points where (a) high_flag = TRUE AND (b) at least one other background site in the urban area had a valid MDA8 O_3 or daily average $PM_{2.5}$ value for that day AND (c) the valid values at the other background sites were all more than 10% larger than the preliminary background estimate. Note that the latter two criteria have to be true for replacing the preliminary background estimate with a value from a different background site to make a significant impact on any subsequent analysis. Data points that met all three criteria are flagged in the csv files in a column called "final_flag", with a value of TRUE meaning that the above criteria were satisfied. The number of points with final_flag = TRUE for each urban area is shown in Table B.3. For these points, we have included the AQS site number and the MDA8 O_3 or daily average $PM_{2.5}$ value for the background site with the second largest value in the csv files as an alternate background estimate.

However, we only replaced the preliminary background value if:

1. The final_flag = TRUE
2. The estimate was for the HGB or BPA areas, as these areas near the Gulf of Mexico could plausibly have times when the gulf/lake breeze front affects some of the outlying background sites, but does not affect the urban area as a whole.
3. The preliminary background site was between the city and the Gulf of Mexico (or the city and Sabine Lake). These sites are given in Table B.4.

These final estimates were delivered to TCEQ as part of Deliverable 3.1. The files in this deliverable are discussed in Section AB.2.3.

3.2 Temporal Trends of Background O_3 and $PM_{2.5}$

Figure 7 and Figure 8 show the seasonal (left) and annual (right) trends in the background MDA8 O_3 for the six urban areas of interest, while Figure 9 and Figure 10 show the same for background daily average $PM_{2.5}$.

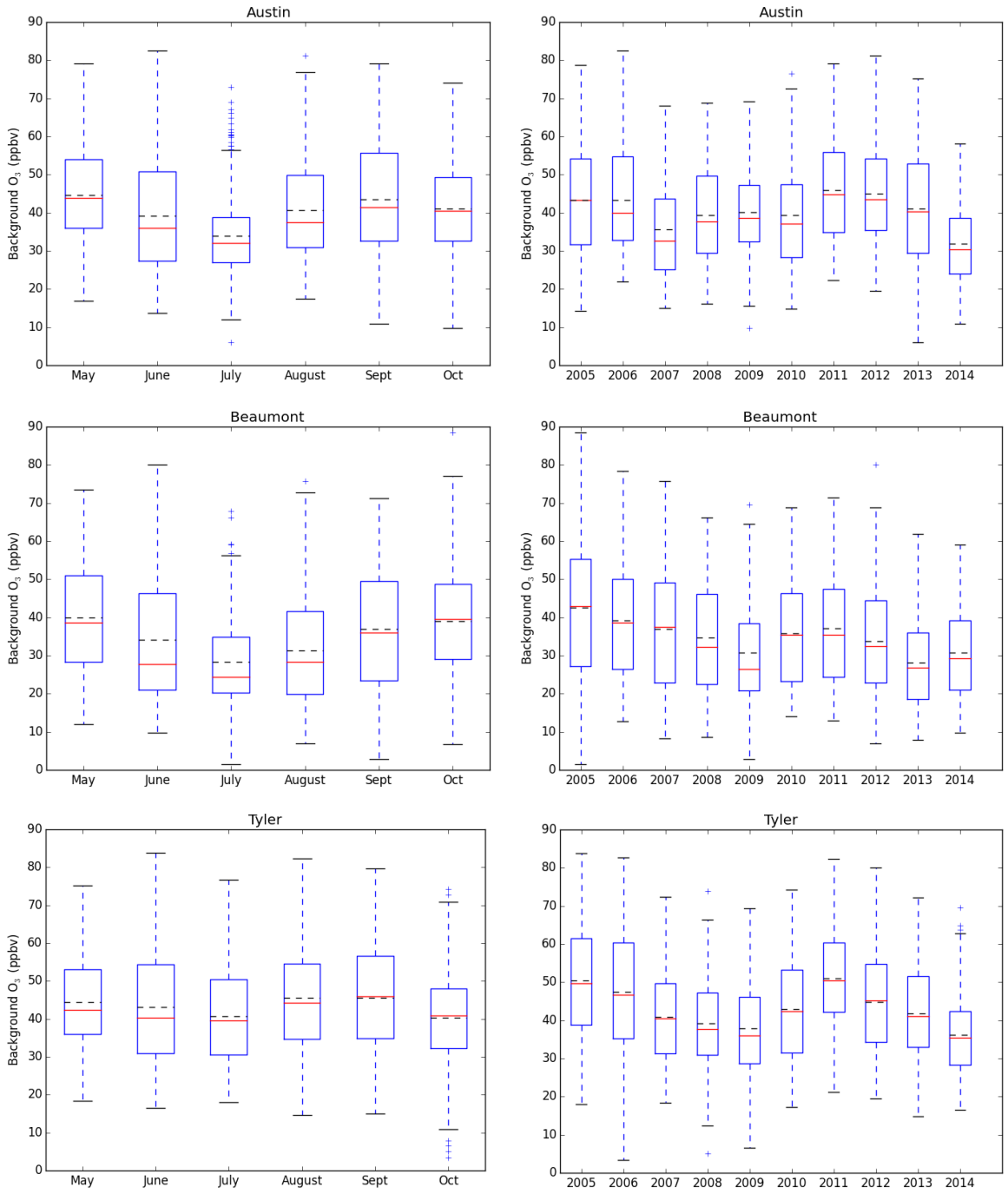


Figure 7. Box-and-whisker plots for the background MDA8 O_3 for Austin/Round Rock, Beaumont/Port Arthur and Tyler-Longview-Marshall as estimated using the TCEQ method. The red line is the median, the dashed black line is the mean. Box edges show the 25th and 75th percentiles (Inter-Quartile Range, or IQR), the whiskers show the data range up to ±1.5*IQR and the crosses show the outliers beyond 1.5*IQR.

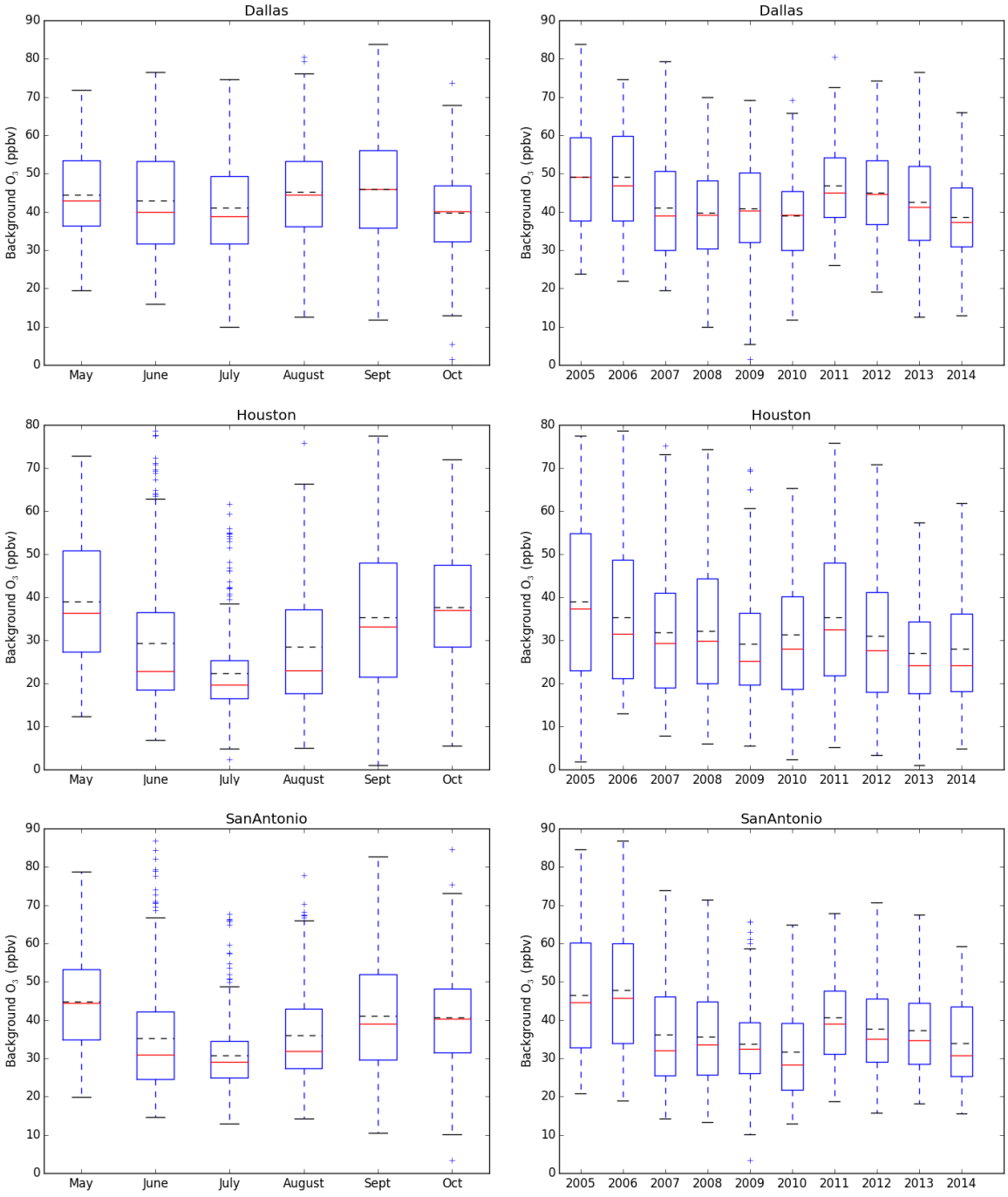


Figure 8. Box-and-whisker plots for the background MDA8 O_3 for Dallas/Fort Worth, Houston/Galveston/Brazoria, and San Antonio as estimated using the TCEQ method.

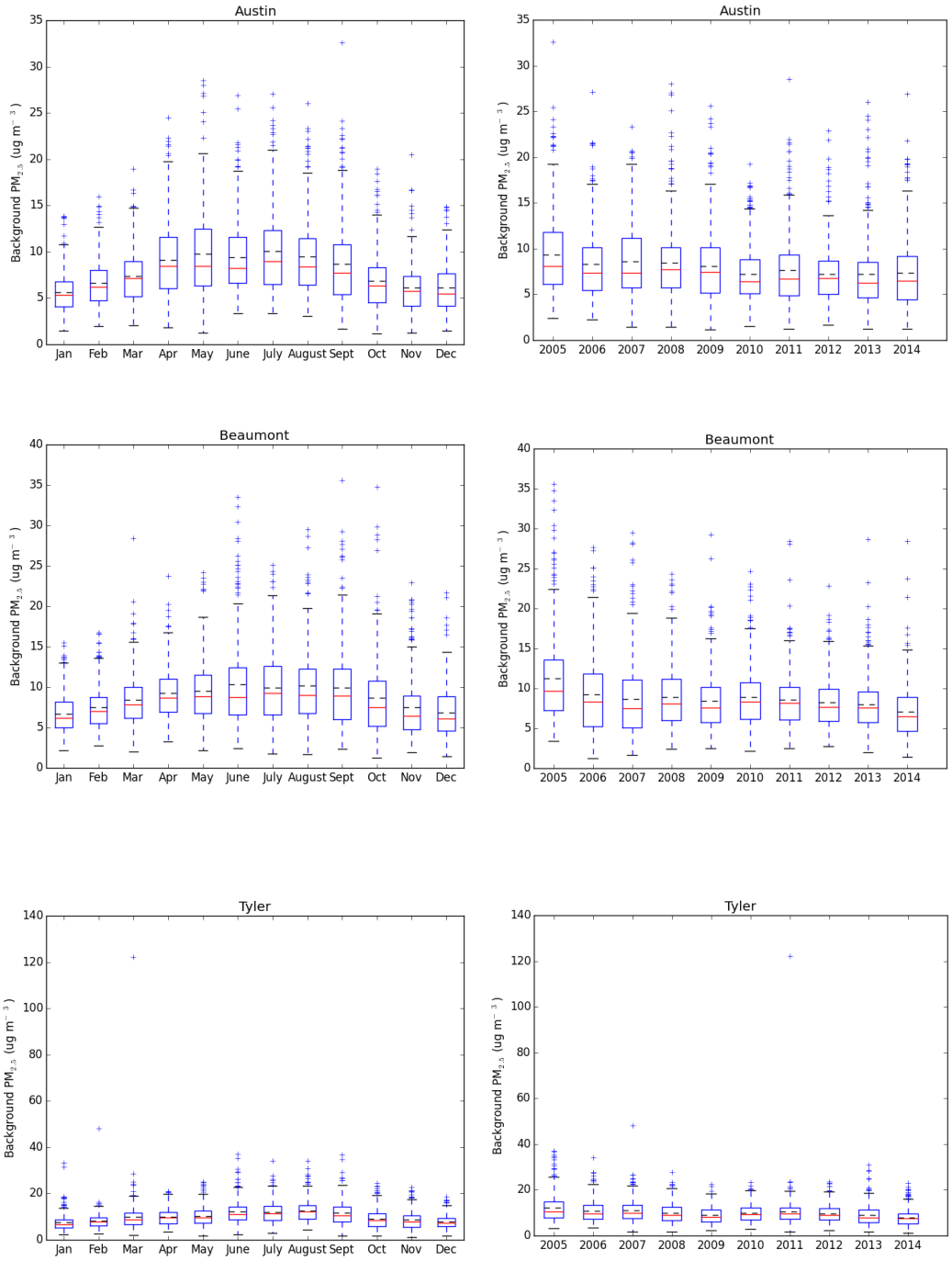


Figure 9. Box-and-whisker plots for the background daily average $PM_{2.5}$ for Austin/Round Rock, Beaumont/Port Arthur and Tyler-Longview-Marshall as estimated using the TCEQ method.

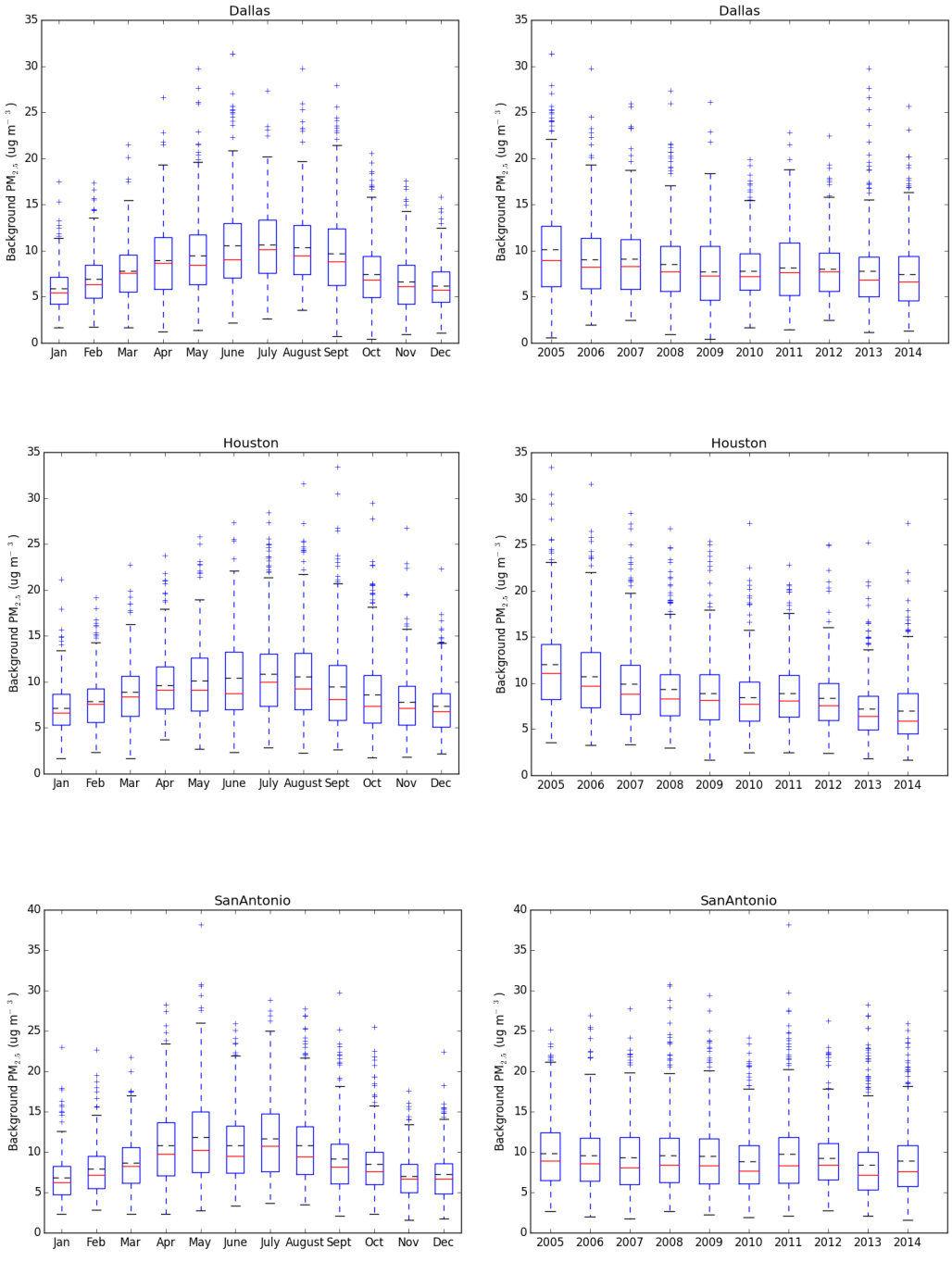


Figure 10. Box-and-whisker plots for the background daily average $PM_{2.5}$ for Dallas/Fort Worth, Houston/Galveston/Brazoria, and San Antonio as estimated using the TCEQ method.

Background MDA8 O_3 is fairly constant with month during the O_3 season for TLM and DFW, but has a July minimum for the other urban areas. In contrast, median background $PM_{2.5}$ peaks in June and July. HGB and BPA show a clear decreasing trend in background O_3 and $PM_{2.5}$ – the analysis of Section 2.4 suggests that these trends are reduced but still present when the effect of varying meteorology is accounted for. The range of values for a given month or year is large for all cities, with HGB having the largest O_3 spread and most points outside 1.5 times the Inter-Quartile Range (IQR), possibly due to the fact that it has the largest dataset (however, note that these extreme points have been kept in all of the analyses discussed in this report). The month of June for O_3 in particular has a large spread of values, including occasional very high (greater than 70 ppbv) background values for HGB and SA. However, it should be noted that the frequency of "high" (greater than 55 ppbv) background O_3 days in June is only 11% and 12% in June for HGB and SA, respectively, in contrast to May, which has larger frequencies (17% and 22%, respectively). In both urban areas, the frequency of "high" background O_3 seems to have a strong inverse relationship with the frequency of flow from the Gulf (synoptic map type 2, see Section 4.1.1 and Figure 20), as expected.

3.3 Alternative Methods To Determine Regional Background O_3 and $PM_{2.5}$

3.3.1 Determining Background O_3 with PCA

In Langford et al. (2009), principal component analysis (PCA) was applied to the large dataset of MDA8 values at 30 sites across HGB for a 2.5-month timespan (August to October 2006). The PCA approach attempts to isolate the large day-to-day regional changes in the MDA8 O_3, and Langford et al. (2009) were able to associate regional meteorological patterns with the patterns of covariance as determined by the PCA using associated meteorological data provided by the National Centers for Environmental Prediction (NCEP) reanalysis. They found that nearly 84% of the variance in the MDA8 ozone near HGB was described by the first Principal Component (PC1) and could be attributed to the regional background ozone concentration. PC2 and PC3 described 6% and 3.5% of the variance in MDA8 ozone, and were attributed to local photochemistry and transport, respectively. After determining that PC1 described that large majority of variance and represented regional background ozone, the following equation was applied to calculate the hourly background ozone for HGB, $O_3^{RM}(t)$,

$$O_3^{RM}(t) = \overline{O^3} + \sigma(O^3)\mathfrak{f}_1\alpha_1(t)$$

where $\overline{O^3}$ is the mean of all MDA8 ozone values for the entire time period, $\sigma(O^3)$ is the standard deviation of that mean, $\mathfrak{f}_1$ is the variance contribution of PC1 (0.84 in Langford et al., 2009) and $\alpha_1(t)$ is the score (or amplitude) of PC1 at each hour.

Performing a PCA of our MDA8 O_3 data for the Group 1 urban areas required that we first create a full, interpolated dataset without any missing values. We calculated MDA8 values following the steps described in Appendix B. We then filtered out any sites where less than 75% of data points were valid for the 10-year period during the ozone season (May to October, 2005-2014). Next, we spatially interpolated the dataset to replace any missing MDA8 O_3 values. If the data point for that day was located outside of the cluster of sites with valid data points, we applied a nearest-neighbor interpolation, and if that point was located within the cluster of sites with valid data points, we applied a cubic interpolation in latitude and longitude.

Once this complete dataset was established, we applied the PCA using the eigenvector-eigenvalue calculation in R to the entire 10-year time span for HGB, which resulted in a similar variance contribution to that of Langford et al. (2009), where PC1, PC2 and PC3 had variance

contributions of 83%, 6%, and 2%, respectively. However, when assuming PC1 represented the regional background contribution and applying the above equation, the values were well correlated to our original background estimates from the TCEQ method (see Appendix B), but produced a much larger and unrealistic range of background concentrations (-100 to 250 ppbv), as seen in Figure 11.

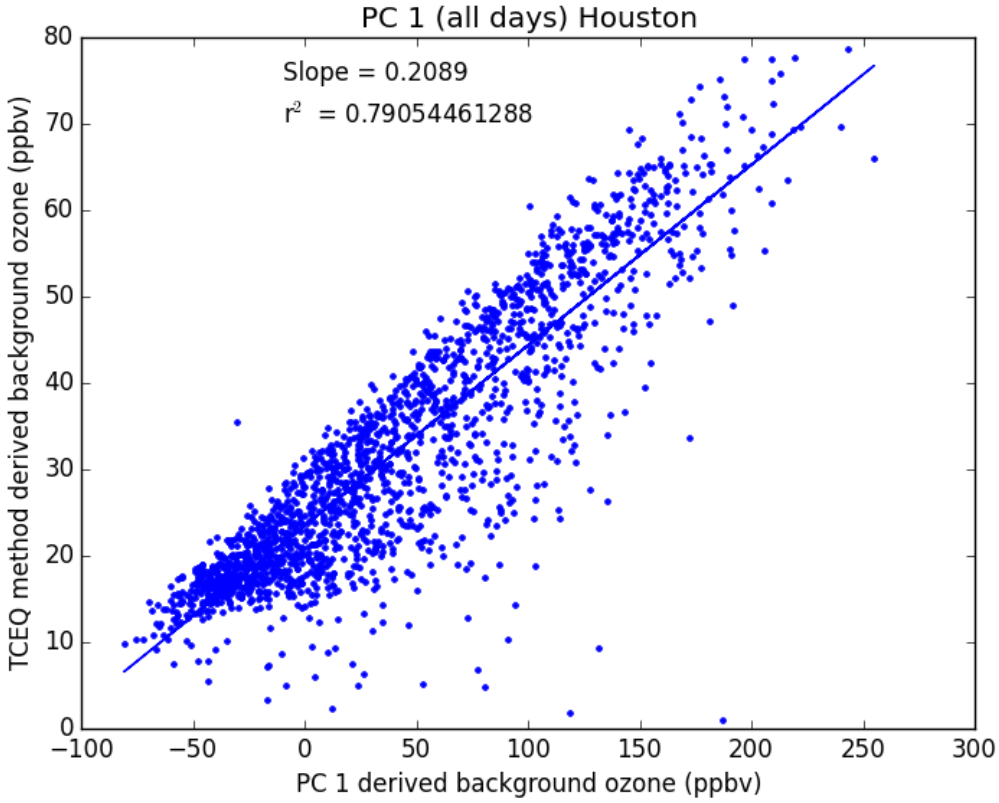


Figure 11. PCA-derived background ozone in Houston/Galveston/Brazoria applied over the entire ozone season dataset (x-axis, 10 years and all sites) compared to our original TCEQ method of determining background ozone (y-axis).

After further discussion via email with Langford, and examining the day-of-year functional fits from our meteorological analysis of MDA8 O_3 described in Appendix A. Effects of Meteorology on O_3 and $PM_{2.5}$ Trends, we found that the mid-ozone-season wind shift in July may be contributing to the unrealistic PC1 scores and thus giving unrealistic background concentrations. July is the peak month in terms of the frequency of synoptic flow from the Gulf (MT 2, see Section 4.1), and the months following this period generally have weak synoptic forcing associated with stagnant conditions (MT -999). In addition, the Gulf flow in July may give significantly different spatial distributions of O_3, as the strength or weakness of the Gulf flow will change which stations have the highest O_3. We thus performed the PCA analysis separately for two periods, May-July and August-October, as the meteorological patterns in each period are expected to be similar. The results are presented in Figure 12, where the correlation with the background from the TCEQ method remains high, and the range of concentrations is more realistic.

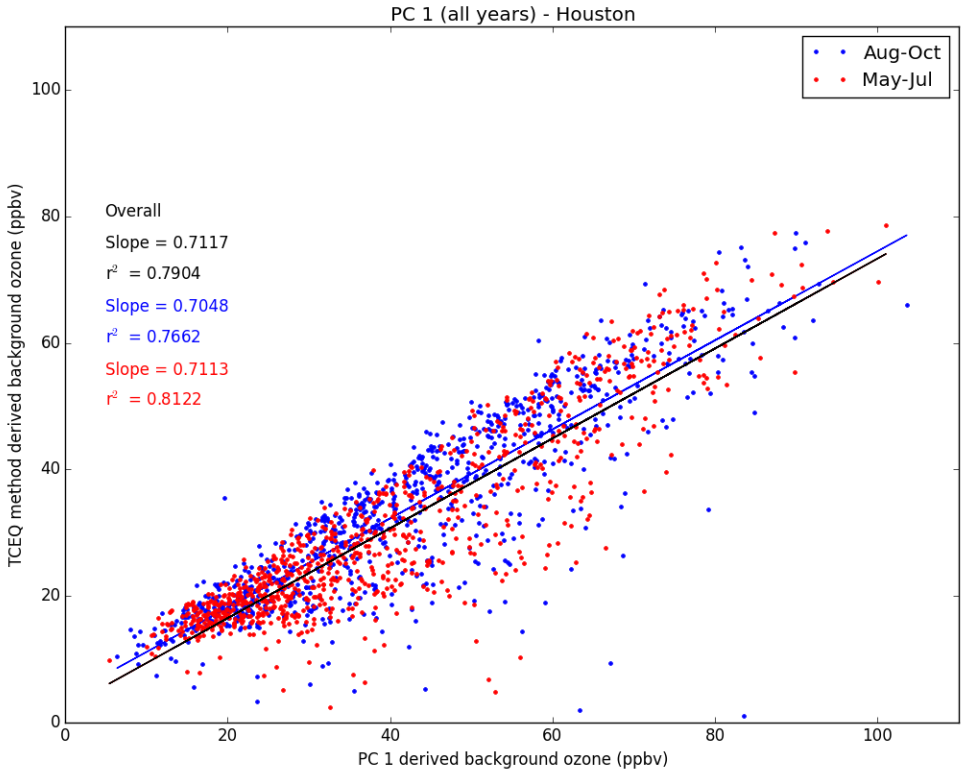


Figure 12. PCA-derived background ozone in Houston/Galveston/Brazoria compared to the original TCEQ method. This approach applied the PCA over time spans during the ozone season; May to July (red) and August to October (blue) with the overall slope and r^2 value printed in black.

Similar correlation results were seen for the remaining Group 1 urban areas, ARR, DFW and SA where the r^2 values were 0.91, 0.89, 0.88, respectively. However the slopes indicate differences between the background ozone estimated by the TCEQ method and those estimated by PCA, and the relationship between the two background estimates varies significantly between cities, which differ between the four urban areas examined (i.e., 0.71, 1.5, 0.75, and 1.27 for HGB, ARR, DFW, and SA, respectively). Further work is needed to determine the reason for the differences in these two estimates of background O_3, and why the differences vary with urban area.

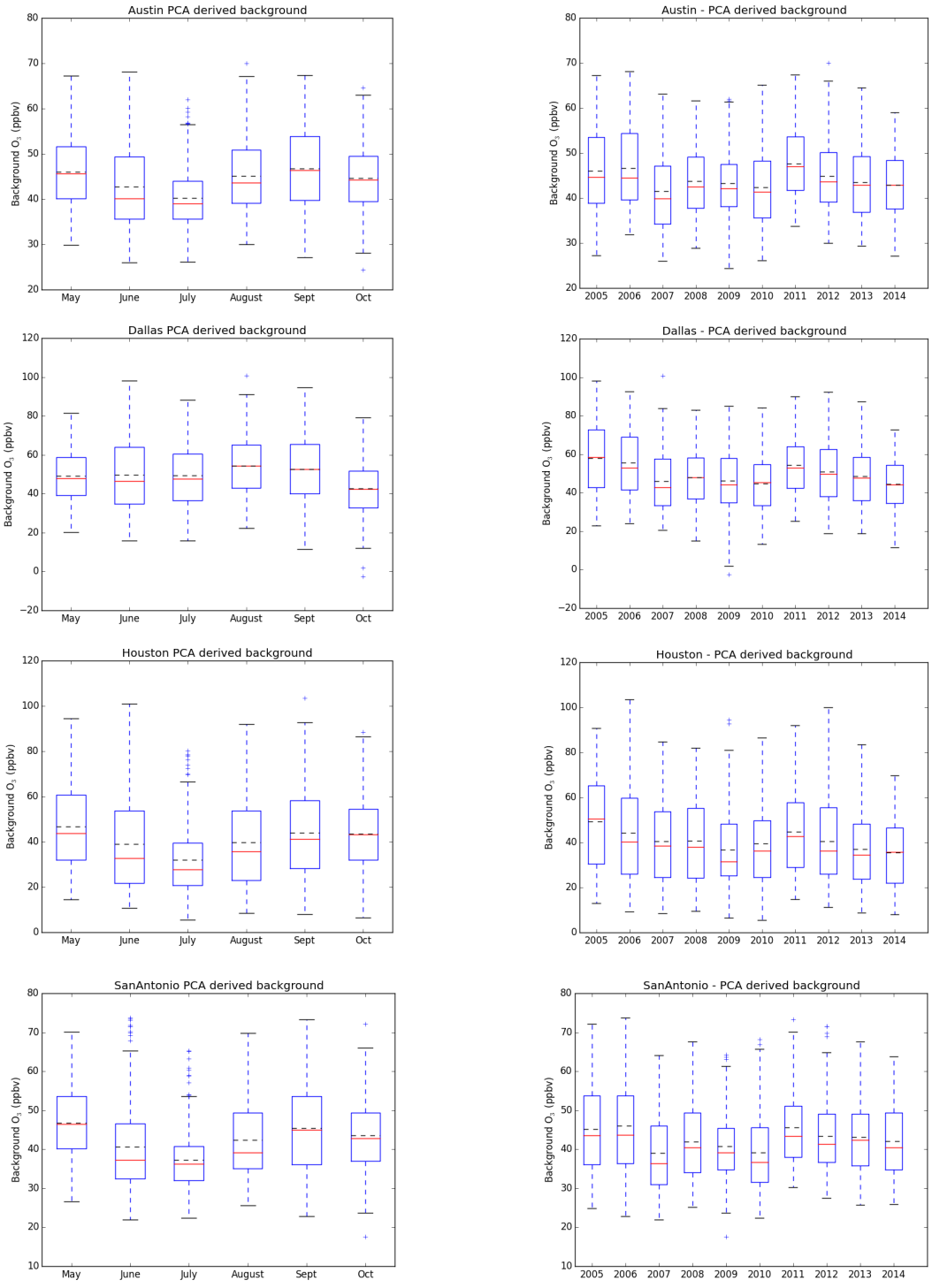


Figure 13. Monthly (left) and yearly (right) background ozone (ppbv) as derived using the PCA method. Box plots during the ozone season for Austin/Round Rock, Dallas/Fort Worth, Houston/Galveston/Brazoria and San Antonio are shown.

3.3.1.1 PCA-derived Background O_3 Temporal and Spatial Analysis

Figure 13 shows the monthly and yearly background MDA8 O_3 box plots for the Group 1 urban areas as determined by the PCA analysis described above. These results compare well with the temporal analysis discussed above in Section 3.1 of MDA8 background ozone as determined using the TCEQ method. The seasonal background ozone (left) peaks in September and has a minimum concentration in July. This is consistent with the GAM analysis (see Section A.1.5.2) and the TCEQ method (see Section 3.2) for all of the urban areas except for DFW, where there is a less pronounced trough, and the minimum average background occurs in October. The yearly background MDA8 ozone also appears to be similar for all urban areas and to the July progress report results (however less pronounced), where there is a slight overall average decrease but with year-to-year variation.

3.3.1.2 Comparing Trends in Background O_3 from the PCA and TCEQ methods

After completing background estimates using the PCA and TCEQ methods, we performed a comparison of the yearly average concentrations for both methods (Figure 14). As in Figure 12, the PCA-method background estimates are larger than those of the TCEQ method for all urban areas. However, both methods indicate a decrease in background O_3 concentrations from 2005 to 2014, consistent with the Houston results presented in Berlin et al. (2013). The slopes, trends and decreases are similar for each method in all urban areas except for SA where the slopes are -0.1 and -1.0 for the PCA and TCEQ method, respectively.

3.3.2 Determining Background $PM_{2.5}$ with PCA

The same PCA approach was applied to the $PM_{2.5}$ dataset provided by TCEQ. First, a complete dataset was established, with the same interpolation method described above. Then the PCA was applied to the entire 10-year timespan, where the entire year was analyzed, not just the ozone season. The initial variance contributions in HGB for PC1, PC2 and PC3 were 87%, 4.4% and 3.0%, respectively, suggesting that PC1, if assumed to be associated with the regional background, plays an even more significant role to the overall variance than it did for O_3. However, similar to our background ozone PCA method, when we applied the equation for Langford et al. (2009) to calculate the regional background $PM_{2.5}$ (Figure 11) the correlation was reasonable, but the range of concentrations proved unrealistic (-10 to 60 μg m^3, see Figure 15).

After further investigation and reference to the GAM meteorological analysis from Appendix A. Effects of Meteorology on O_3 and $PM_{2.5}$ Trends,we recognized that there could be more than one potential meteorological shift throughout the year influencing $PM_{2.5}$., as well as other factors. After attempting to split the years up into different, meteorologically-similar periods, we found that even if the PCA was applied to each month individually, the Langford et al. (2009) approach still gave an unphysical range of estimated background concentrations. For example, Figure 16 shows that even when PCA is only applied to the month of July for 2005-2014, the background estimates from Equation 1 are still unphysical.

We attempted a variety of subsections in the year, including the example in Figure 17, which is split up into; 1) April-July, 2) August-October and 3) November-March. However, none of these attempts resulted in significant improvements, thus further analysis is needed to derive a more comprehensive PCA-based method for determining background $PM_{2.5}$ for the HGB area. For the other 3 urban areas that we applied the PCA method to, similar unrealistic results were seen. Thus we conclude that this PCA is not a reasonable way to derive background estimates for $PM_{2.5}$ for the four urban areas considered here.

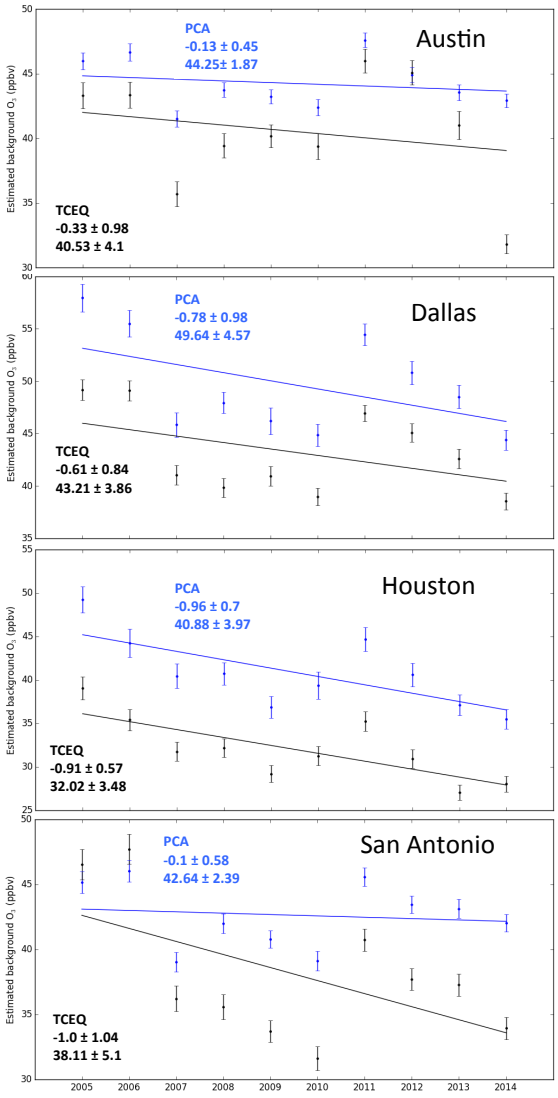


Figure 14. Comparing the yearly average estimated background O_3 using the PCA method (blue) and TCEQ method (black) for each of the Group 1 Urban areas. The first line of text gives the trend (ppbv/year) and the 95th confidence interval of the trend, while the second line is the mean and standard deviation of the annual averages. The error bars represent one standard error from the mean for each year.

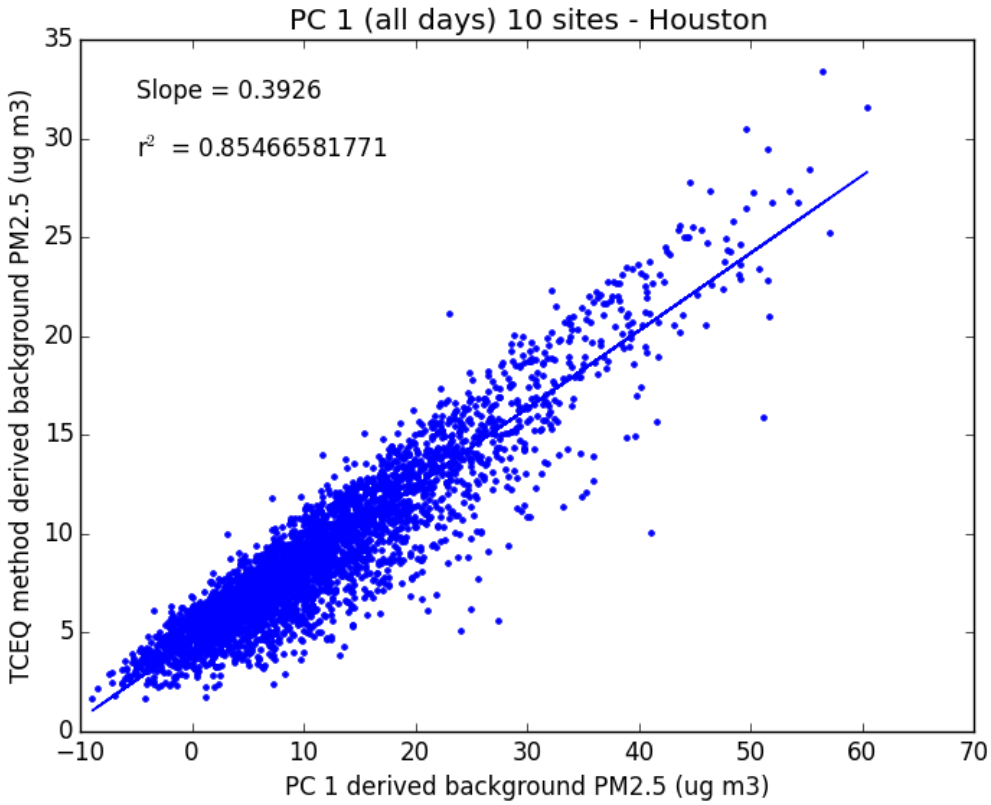


Figure 15. PCA-derived background $PM_{2.5}$ in Houston/Galveston/Brazoria compared to the original TCEQ method. The PCA was applied to the entire 10-year time span for all sites.

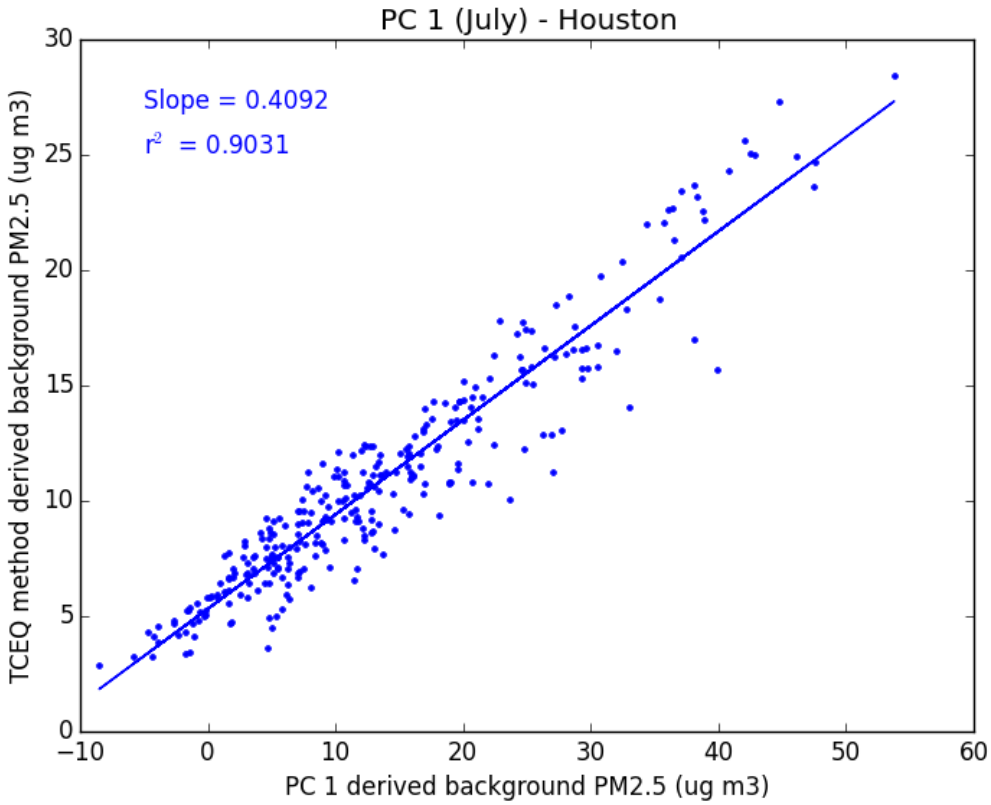


Figure 16. PCA-derived background $PM_{2.5}$ in Houston/Galveston/Brazoria in just July months compared to the original TCEQ method.

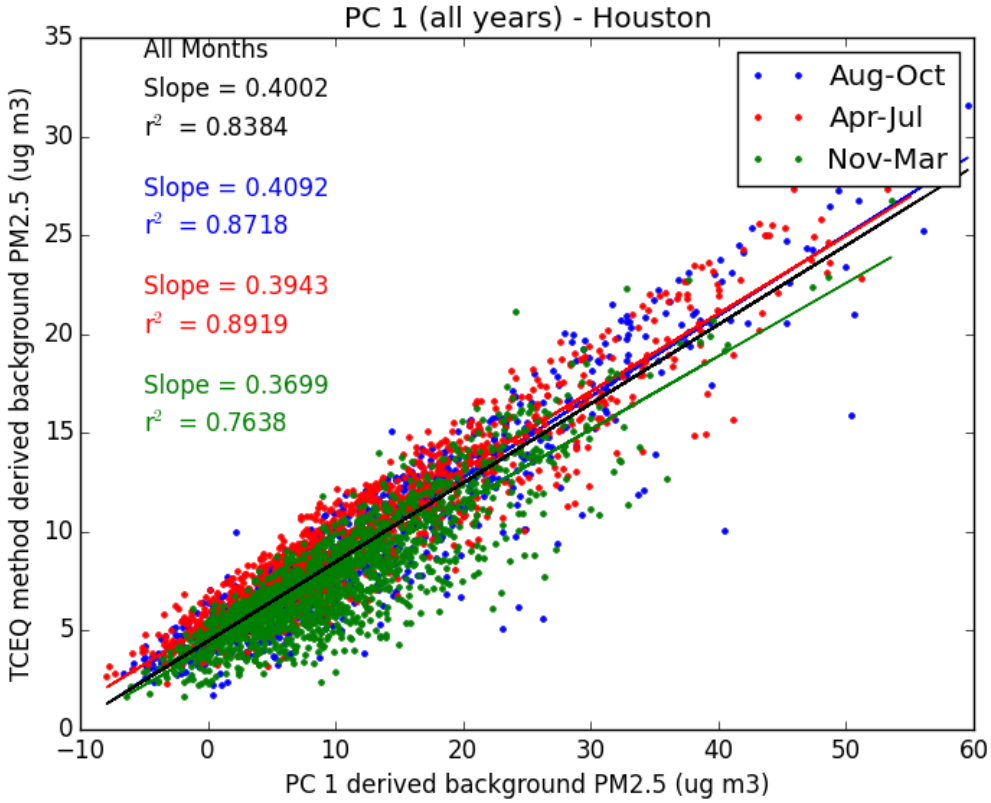


Figure 17. PCA-derived background $PM_{2.5}$ in Houston/Galveston/Brazoria compared to the original TCEQ method. The PCA was applied to 3 different time spans: August to October (blue), April to July (red) and November-March (green).

3.3.3 Determining Background O_3 with Satellites

We performed a literature review in order to explore alternative methods to measure surface O_3 that could improve the understanding of spatial and temporal ozone trends across Texas. Such a dataset could improve our understanding of local and background ozone trends and the contributing photochemical and meteorological conditions leading to high ozone events.

Zoogman et al. (2014) discuss a data assimilation system that uses the GEOS-Chem chemical transport model (CTM) and the proposed design for the GEO-CAPE (GEOstationary Coastal and Air Pollution Events) mission to calculate a better representation of surface ozone. They argue that the assimilation of satellite ozone data into a CTM can be further improved by using correlated multispectral CO measurements, which have better boundary layer sensitivity than ozone. The observed model-transport error of CO could assist in identifying model-transport ozone errors, resulting in improved surface ozone representation. They apply this framework to a regional-scale Observing System Simulation Experiment (OSSE) of GEO-CAPE over North America and conclude that this technique could provide improved constraints on surface ozone. Additionally, a satellite that is more sensitive to O_3 in the boundary layer but less sensitive than CO in the boundary layer would also improve the results. However, GEO-CAPE does not yet have a launch date scheduled, and existing satellite instruments have insufficient coverage and vertical sensitivity to reliably separate boundary-layer O_3 from the values in the free troposphere. However, Fu et al. (2013) present an alternate technique for tropospheric O_3 detection through

combining the Aura Tropospheric Emission Spectrometer (TES) and Ozone Monitoring Instrument (OMI) retrievals. TES measures radiances in the thermal infrared spectrum and OMI measures in the ultraviolet-visible spectrum. By comparing results to in situ ozonesonde measurements Fu et al. (2013) concluded that the combined instruments have more sensitivity to boundary layer O_3 than each instrument on their own or previous techniques used in the past. Since both instruments are aboard the polar-orbiting Aura satellite, there are two retrievals over Texas each day (at about 1:30 am and 1:30 pm local time). This technique and ozone coverage could allow for better understanding of seasonal ozone trends in the Texas area.

3.3.4 Determining Background $PM_{2.5}$ with Satellites

In contrast, a satellite-based technique to retrieve regional $PM_{2.5}$ concentrations at fine spatial and temporal resolutions is much further developed and shows significant promise. Van Donkelaar et al. (2013) present an optimal estimation algorithm using aerosol optical depth (AOD) from the NASA Moderate Resolution Imaging Spectroradiometer (MODIS) to determine $PM_{2.5}$ concentrations near the surface at a 0.01° latitude x 0.01° longitude resolution. GEOS-Chem is used to provide the a priori values and the AOD/$PM_{2.5}$ relationship, which is a function of the atmospheric vertical structure, aerosol type, and meteorological factors. The LIDORT (LInearized Discrete Ordinate Radiative Transfer model) is used to simulate the dependence of the top of the atmosphere reflectance on the aerosol type and distribution, and AERONET (AErosol RObotic NETwork) measurements of AOD are used for surface validation. Compared to measurements, this optimal estimation technique proves to be better at predicting surface $PM_{2.5}$ concentrations than the original MODIS and GEOS-Chem outputs, as can been seen in Figure 18. The results from such an approach in Texas could prove to be extremely useful in understanding the spatial and temporal trends of $PM_{2.5}$.

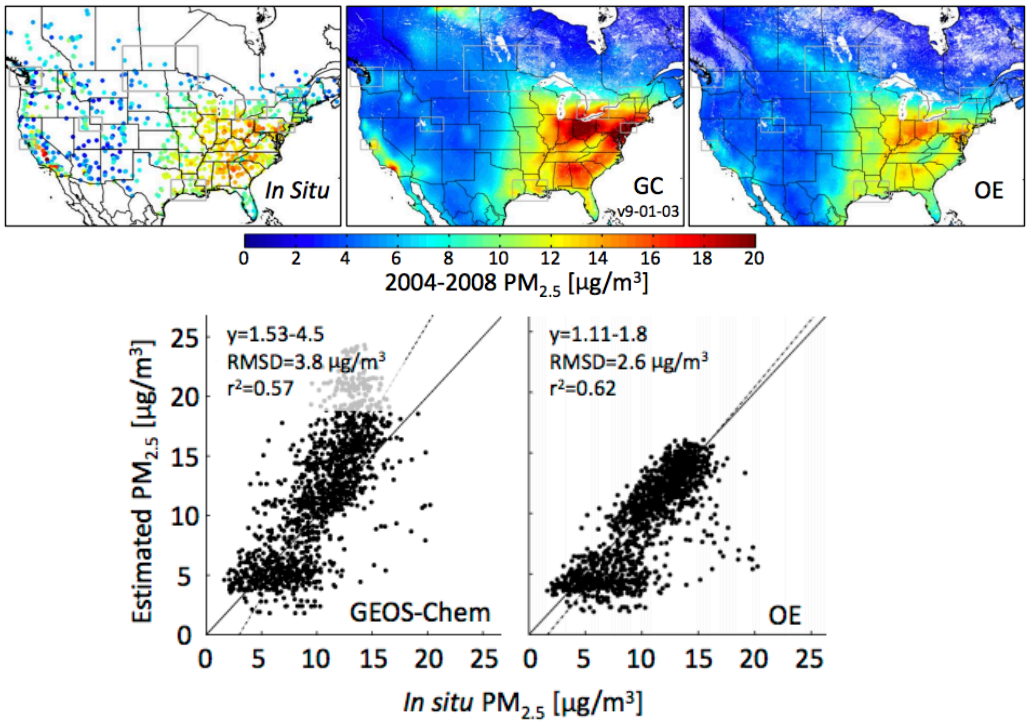


Figure 18. From van Donkelaar et al. (2015). Spatial plot (Top) and scatter plot (bottom) over the US of optimal estimation (OE) approach (far right) for simulating near-surface $PM_{2.5}$ concentrations compared to GEOS-Chem (center) and in situ measurements from the AERONET sites (far left). Presented at the 7th annual GEOS-Chem meeting at Harvard University, 2015.

4 Task 4: Importance of Synoptic/Mesoscale Meteorological Conditions in Explaining/Forecasting Background and Maximum O_3 and $PM_{2.5}$

4.1 Synoptic Map Type Analysis

4.1.1 Technical Method and Results

The synoptic map typing was performed for all days during 2005 – 2014 using the 1200 UTC 850 hPa geopotential height fields from the 32 km North American Regional Reanalysis using the method of Hegarty et al. (2007). The five most common synoptic map types were identified which enabled the classification of 70% of the days during the 10-year study period (and 58% of the days in the ten May-October O_3 seasons) as being under the influence of one of those types. The five types, shown in Figure 19, are described below.

1. MT (Map Type) 1 occurred on 599 days (364 during the O_3 seasons) and featured an anticyclone over the eastern Gulf of Mexico with a trough in the Central Plains extending into northwest Texas. These features produced a general south-southwest flow over much of Texas.

2. MT 2 occurred on 472 days (127 during the O_3 seasons) and featured a cyclonic circulation centered over the Midwest with a ridge extending southeast to northwest over Mexico and extreme southern Texas. This pattern likely produces a light NW flow over much of Texas.

3. MT 3 occurred on 515 days (474 during the O_3 seasons) and featured a large anticyclone centered over the eastern Gulf of Mexico states extending in to the Gulf and westward in to eastern Texas. A broad trough is aligned along the eastern Rocky Mountains. This pattern produces moderate to strong southeasterly flow over much of eastern Texas.

4. MT 4 occurred on 628 days (196 during the O_3 seasons) and featured a broad trough in the Central Plains with an anticyclone centered over the western Caribbean of southern Florida and extending westward in to eastern Texas. This pattern produces a general southwest flow over Texas.

5. MT 5 occurred on 344 days (143 during the O_3 seasons) and features an anticyclone over the western Gulf of Mexico. This pattern features a south to southwestern flow over much of Texas.

The synoptic type classification for each day in the years 2005-2014 are in the file *tceq_map_type.csv* in the final deliverable (see Appendix C). Days that do not fit any of the five types are indicated as type "-999". This generally occurred under conditions of weak synoptic forcing, which is generally consistent with stagnant conditions in the area. Figure 20 shows a chart of the relative frequency of each synoptic type with month. We can see that the frequency of MT 3 (Gulf flow) shows a strong seasonal cycle, peaking in July at ~40% of days from near zero values in the winter. MT 2 and MT 4 shows an opposite seasonal cycle, being much more frequent in winter than in summer. Unclassified days (MT -999) with little synoptic forcing are most frequent in August and September.

We then wrote an R script (*syn_type_boxplot.R*, see Appendix C) to determine the mean, standard deviation, and quartiles of both total and background MDA8 O_3 and daily average $PM_{2.5}$ for each synoptic type. This analysis focused on the Group 1 urban areas from Table 1. Figure 21 through Figure 24 show box plots of the O_3 and $PM_{2.5}$ distributions for each city and synoptic type.

In order to determine if there was a relationship between synoptic type and the likelihood of high total or background MDA8 O_3 and daily average $PM_{2.5}$ values, we first needed a quantitative

definition of a "high" value of each metric. We derived these metrics by examining the 90th percentile of the distribution of each of the four metrics or each of the four Group 1 urban areas, which are shown in Table 5. We then chose criteria that were roughly in line with these 90th percentiles: 70 ppbv for total MDA8 O_3, 55 ppbv for background MDA8 O_3, 17.0 µg m^{-3} for total daily average $PM_{2.5}$, and 13.0 µg m^{-3} for background daily average $PM_{2.5}$.

Table 6 shows the percentage of days below these criteria for each urban area (i.e., the percentile corresponding to the chosen criteria). These criteria values are also shown as horizontal lines on the box plots in Figure 21 through Figure 24. We then used the script *syn_type_boxplot.R* to determine the percentage of days over these criteria for each synoptic type. These values are summarized in Table 7 through Table 10.

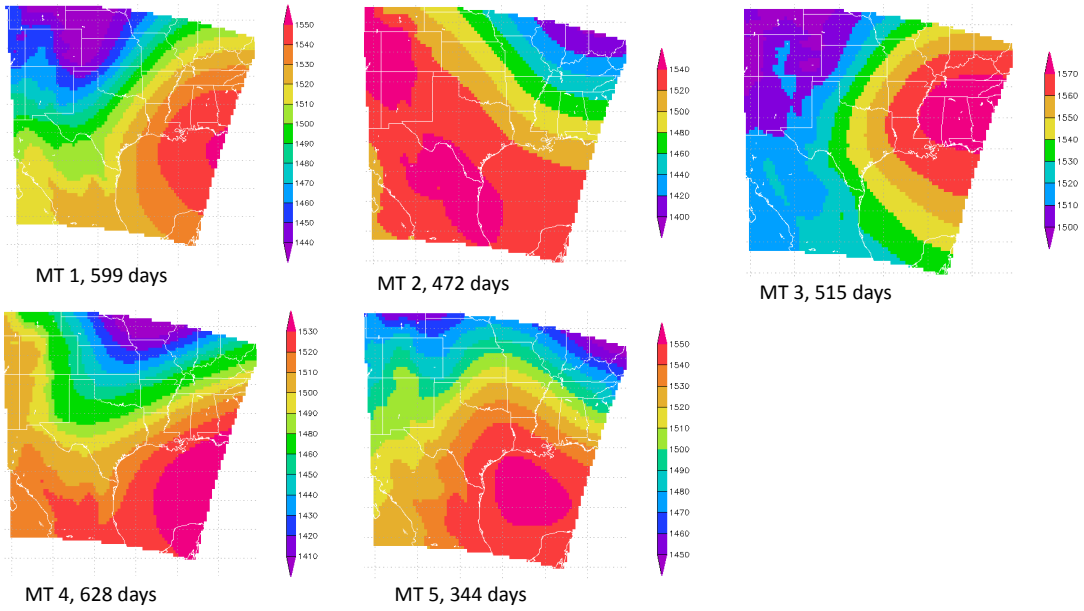


Figure 19. Synoptic maps types determined from 850 mbar geopotential height fields from the 32 km North American Regional Reanalysis using the method of Hegarty et al. (2007).

Table 5. 90th percentile of the total and background (BG) MDA8 O_3 and daily average $PM_{2.5}$ values for each Group 1 urban area for 2005-2010. Only values during the O_3 season (May-Oct.) are considered for O_3.

Urban Area	Total MDA8 O_3 (ppbv)	BG MDA8 O_3 (ppbv)	Total Daily $PM_{2.5}$ (µg m^{-3})	BG Daily $PM_{2.5}$ (µg m^{-3})
HGB	84.3	54.9	20.7	14.6
DFW	83.5	60.6	18.0	13.9
SA	70.3	58.3	17.0	15.6
ARR	67.4	60.1	16.2	13.3

Table 6. Percentage of observations below the criteria chosen to represent "high" values of total and background (BG) MDA8 O_3 and daily average $PM_{2.5}$ values for each Group 1 urban area for 2005-2010. The chosen criteria are in parentheses in the first row. Only values during the O_3 season (May-Oct.) are considered for O_3.

Urban Area	Total MDA8 O_3 (70 ppbv)	BG MDA8 O_3 (55 ppbv)	Total Daily $PM_{2.5}$ (17 µg m^{-3})	BG Daily $PM_{2.5}$ (13 µg m^{-3})
HGB	76.6%	90.1%	77.4%	84.7%
DFW	73.1%	80.2%	87.4%	87.3%
SA	89.7%	86.5%	90.1%	82.5%
ARR	92.9%	82.7%	91.6%	89.3%

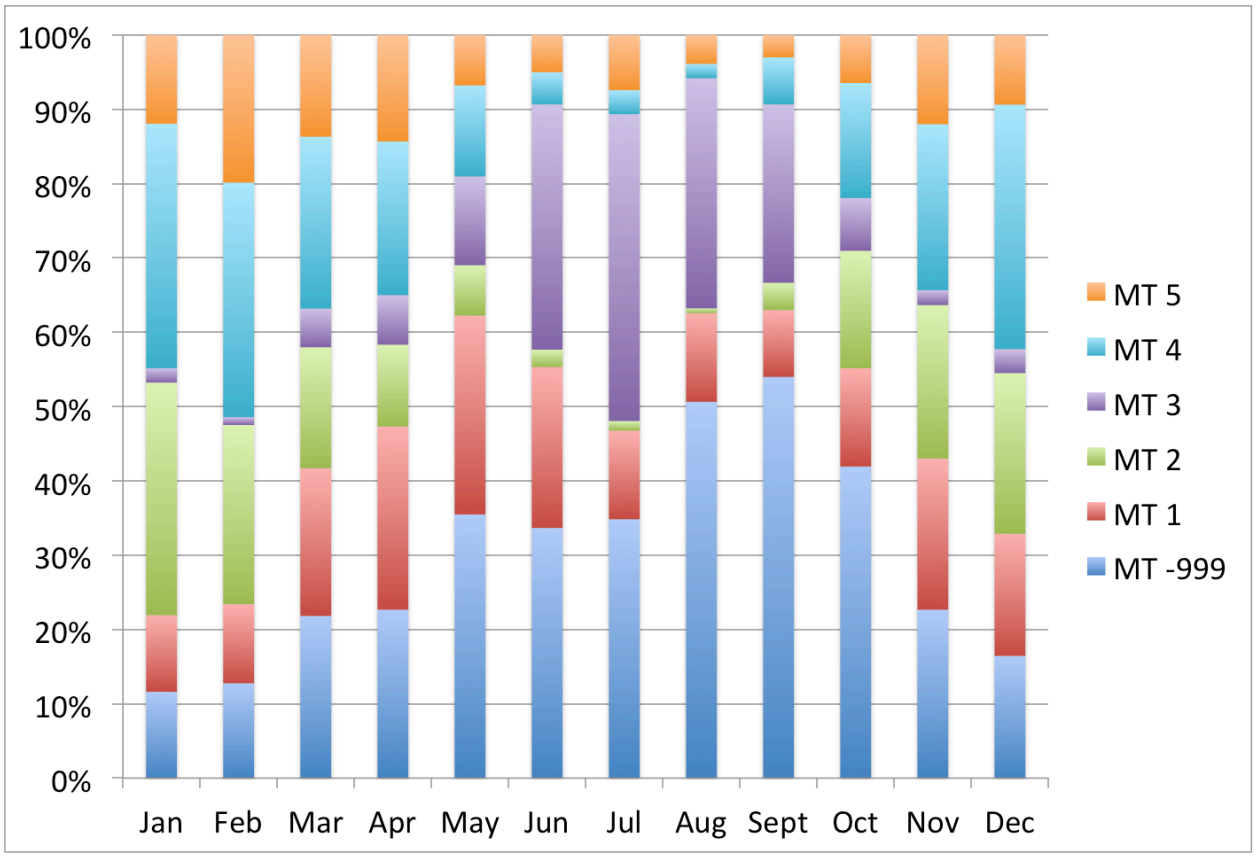


Figure 20. Relative frequency of synoptic map types in each month.

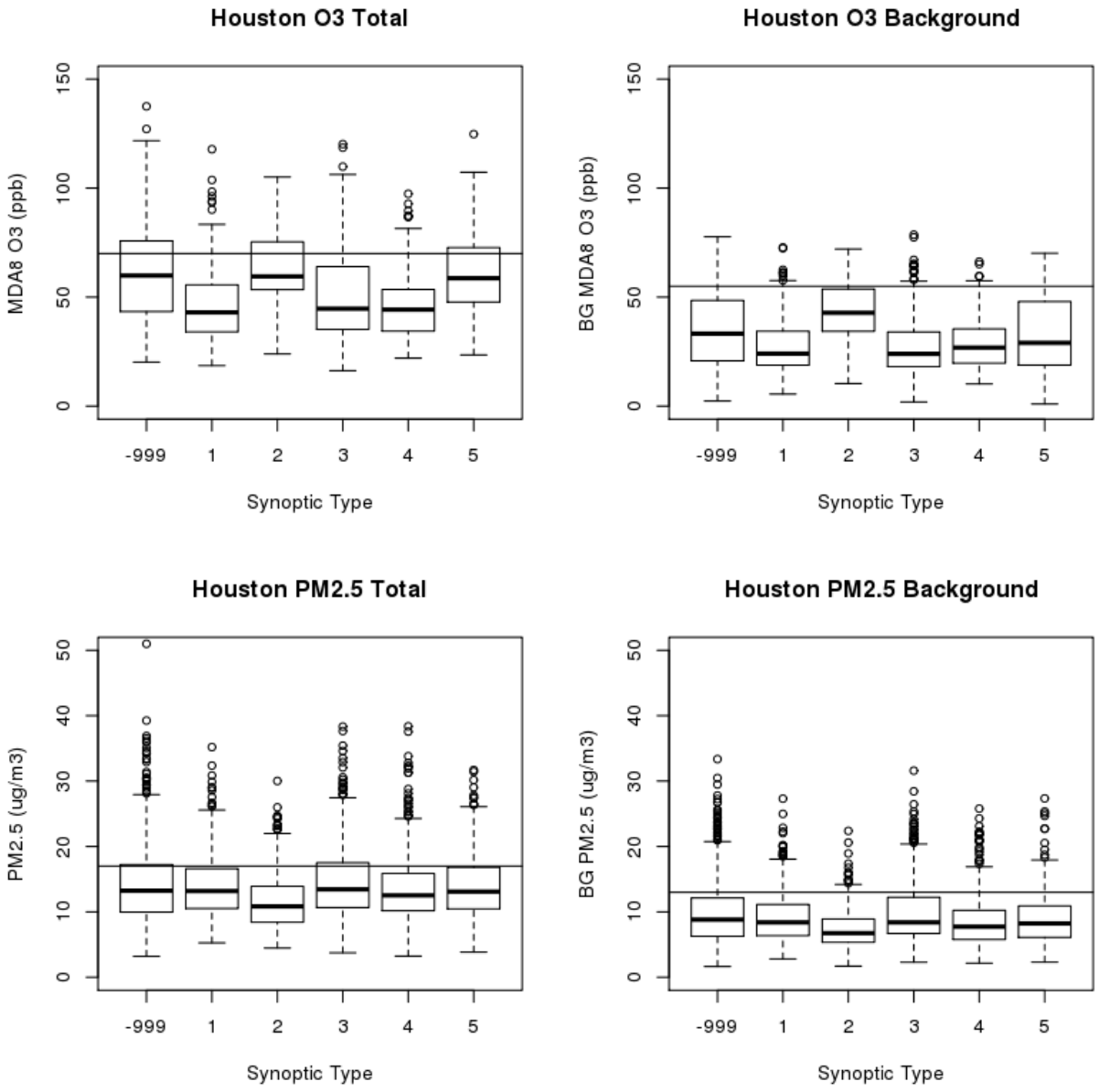


Figure 21. Box and whisker plots of the distributions of total MDA8 O_3 (ppbv, top left), background MDA8 O_3 (ppbv, top right), total daily average $PM_{2.5}$ (µg m^{-3}, bottom left), and background daily average $PM_{2.5}$ (µg m^{-3}, bottom right) for the Houston/Galveston/Brazoria urban area. The thick black line is the median of the distribution, the boundaries of the boxes are the 25th and 75th percentiles, and the whiskers cover the range of the data or all values within 1.5 times of the interquartile range (IQR) of the box, whichever is smaller. The circles denote outliers beyond 1.5 × IQR of the box. The horizontal lines show the criteria denoting "high" values of each metric, as discussed in the text.

Table 7. Percentage of observations above the criteria chosen to represent "high" values of total and background (BG) MDA8 O_3 and daily average $PM_{2.5}$ values for the Houston/Galveston/Brazoria urban area. The chosen criteria are in parentheses in the first column. Only values during the O_3 season (May-Oct.) are considered for O_3.

Pollutant Metric	Synoptic Type	Percentage Above Criteria (%)
Total MDA8 O_3 (70 ppbv)	-999	36.2
	1	8.3
	2	34.0
	3	17.6
	4	11.9
	5	29.0
BG MDA8 O_3 (55 ppbv)	-999	14.2
	1	5.2
	2	20.2
	3	4.2
	4	4.5
	5	15.0
Total Daily $PM_{2.5}$ (17 μg m^{-3})	-999	25.8
	1	22.7
	2	13.7
	3	26.7
	4	19.0
	5	24.7
BG Daily $PM_{2.5}$ (13 μg m^{-3})	-999	19.9
	1	14.7
	2	6.3
	3	20.1
	4	12.0
	5	13.3

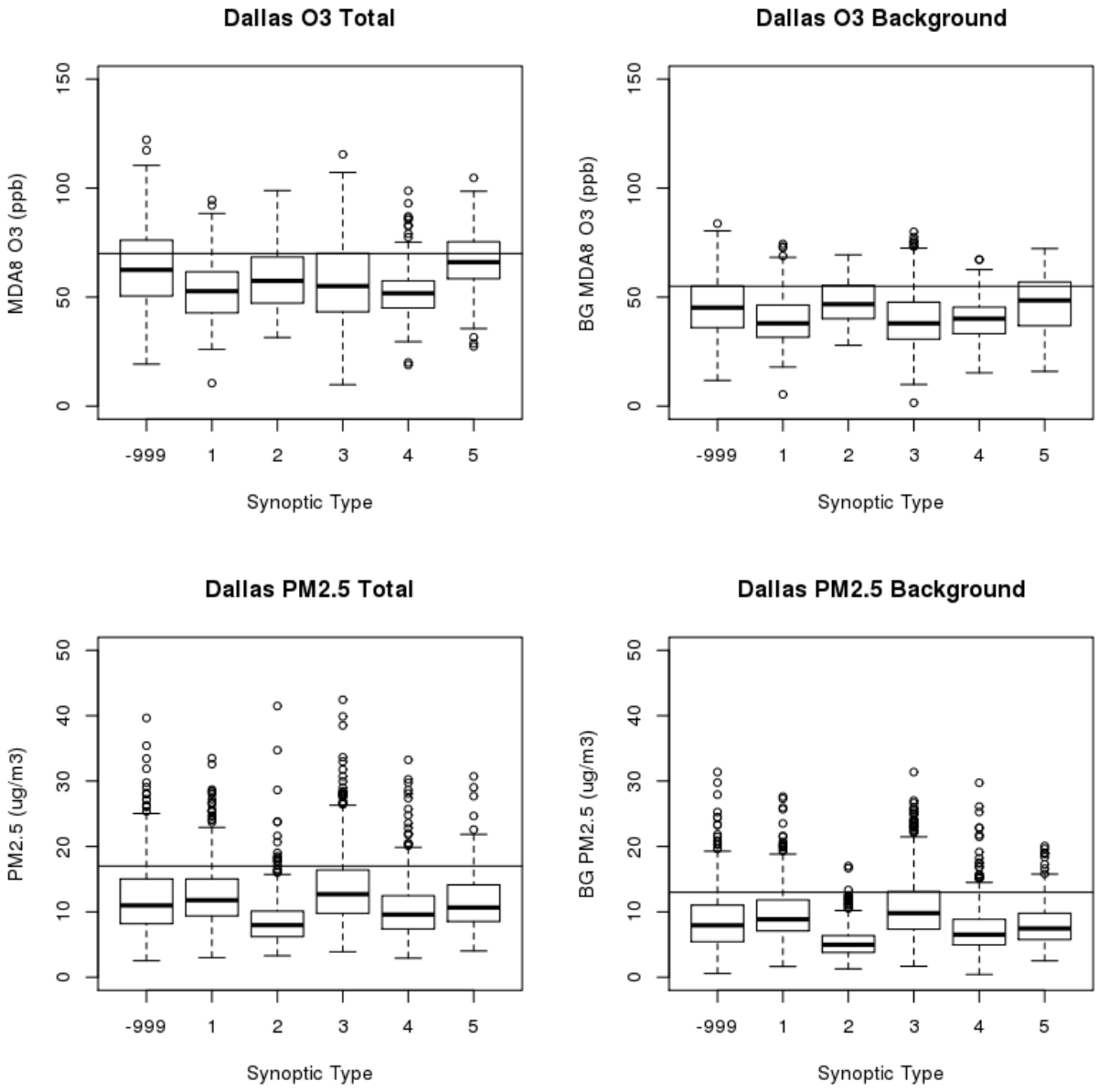


Figure 22. As in Figure 21, but for the Dallas/Fort Worth urban area.

Table 8. As in Table 7 but for the Dallas/Fort Worth urban area.

Pollutant Metric	Synoptic Type	Percentage Above Criteria
Total MDA8 O_3 (70 ppbv)	-999	34.9
	1	11.7
	2	21.3
	3	25.3
	4	13.4
	5	40.0
BG MDA8 O_3 (55 ppbv)	-999	25.2
	1	10.7
	2	26.6
	3	15.6
	4	10.5
	5	30.0
Total Daily $PM_{2.5}$ (17 µg m^{-3})		15.6
	1	14.9
	2	3.0
	3	23.4
	4	6.0
	5	8.2
BG Daily $PM_{2.5}$ (13 µg m^{-3})	-999	14.5
	1	17.4
	2	0.7
	3	25.7
	4	5.7
	5	8.8

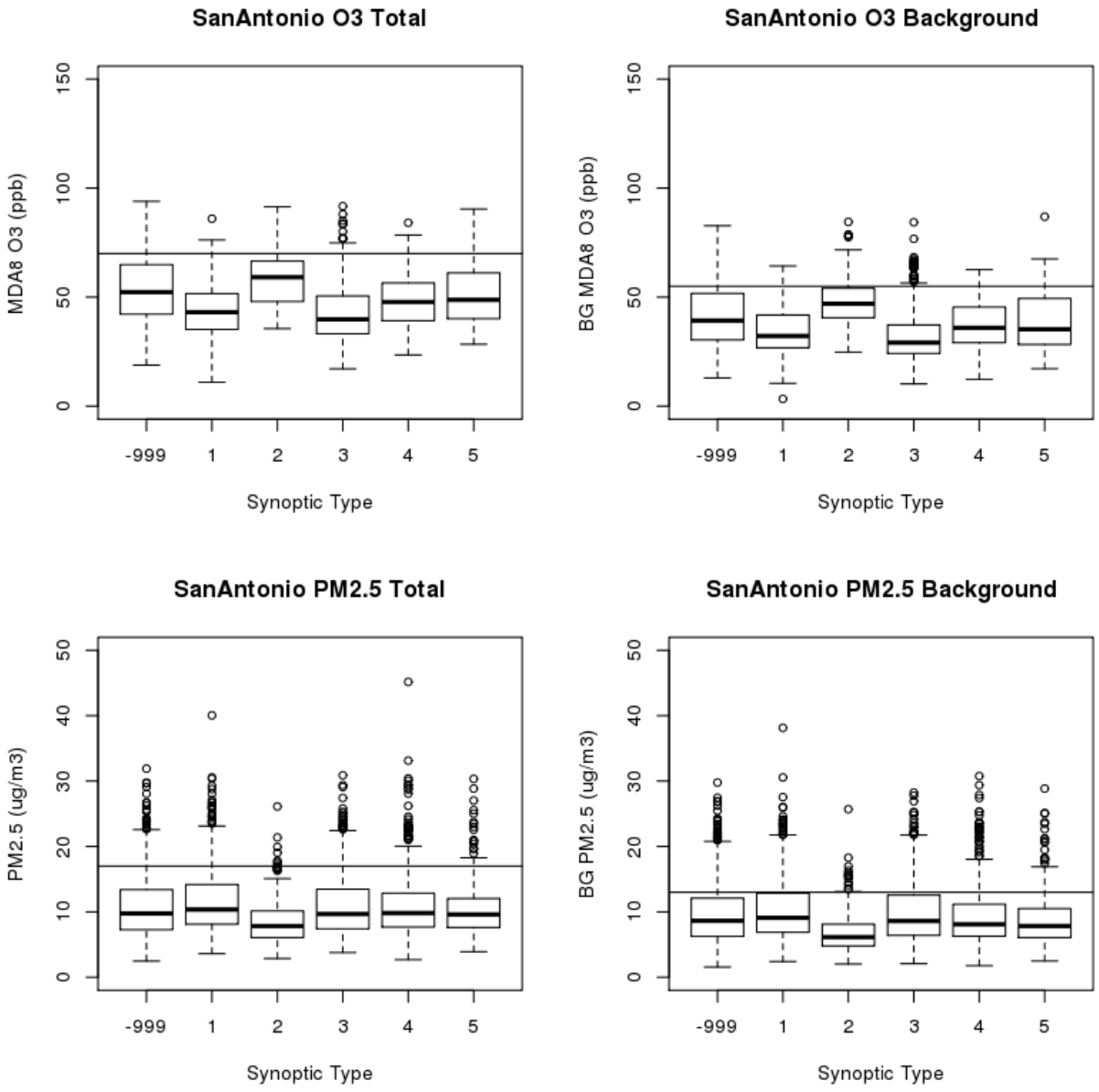


Figure 23. As in Figure 21, but for the San Antonio urban area.

Table 9. As in Table 7 but for the San Antonio urban area.

Pollutant Metric	Synoptic Type	Percentage Above Criteria
Total MDA8 O_3 (70 ppbv)	-999	16.3
	1	2.4
	2	19.2
	3	5.1
	4	5.2
	5	9.0
BG MDA8 O_3 (55 ppbv)	-999	21.1
	1	3.5
	2	22.3
	3	6.4
	4	7.5
	5	15.0
Total Daily $PM_{2.5}$ (17 μg m^{-3})	-999	10.2
	1	14.2
	2	1.6
	3	13.1
	4	10.4
	5	7.1
BG Daily $PM_{2.5}$ (13 μg m^{-3})	-999	19.3
	1	24.5
	2	3.2
	3	21.4
	4	16.8
	5	14.2

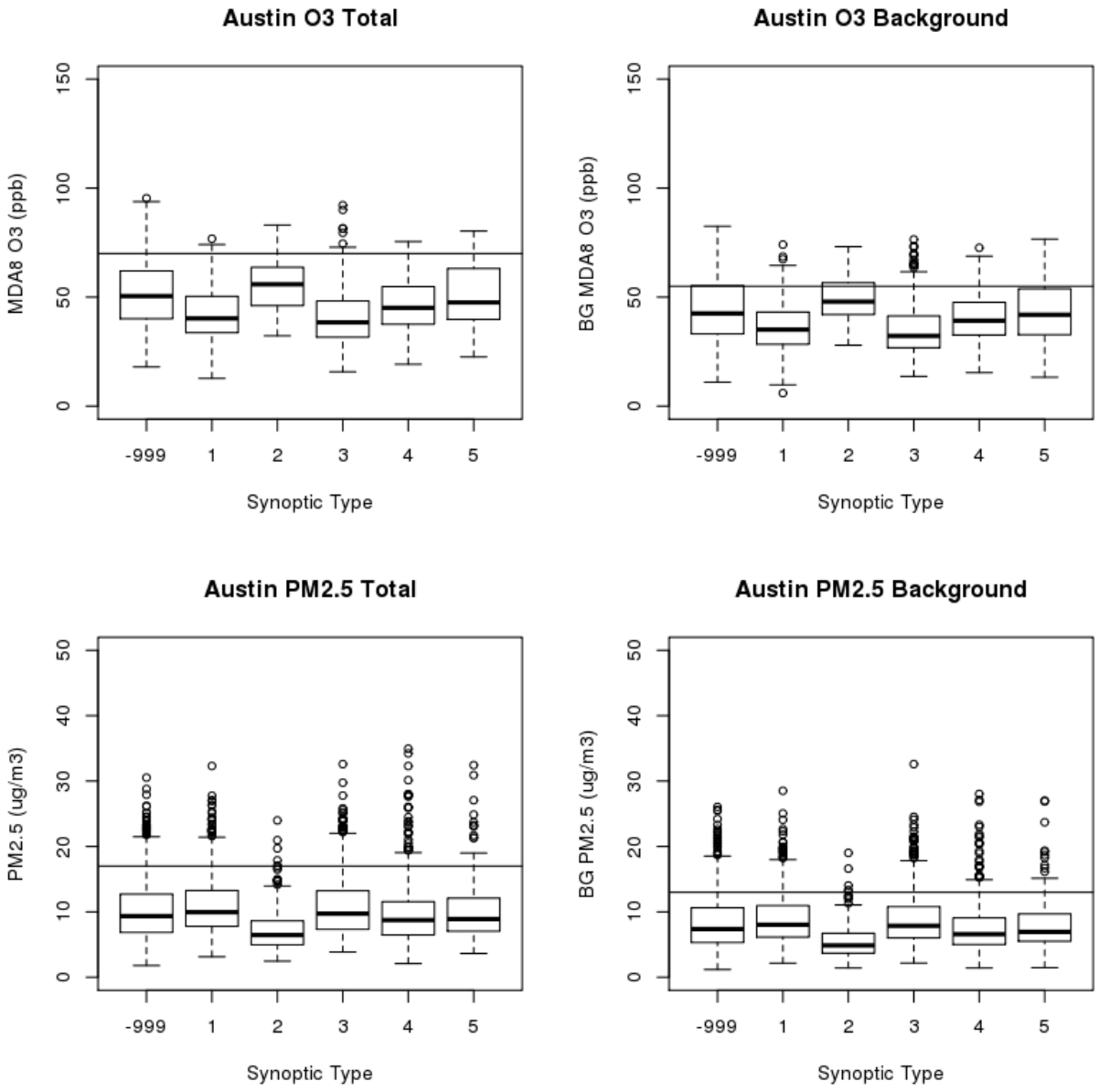


Figure 24. As in Figure 21, but for the Austin/Round Rock urban area.

Table 10. As in Table 7 but for the Austin/Round Rock urban area.

Pollutant Metric	Synoptic Type	Percentage Above Criteria
Total MDA8 O_3 (70 ppbv)	-999	10.7
	1	1.0
	2	9.6
	3	3.7
	4	3.7
	5	14.0
BG MDA8 O_3 (55 ppbv)	-999	25.6
	1	5.2
	2	33.0
	3	8.6
	4	13.4
	5	20.0
Total Daily $PM_{2.5}$ (17 µg m^{-3})	-999	9.3
	1	10.6
	2	1.1
	3	12.8
	4	8.1
	5	5.9
BG Daily $PM_{2.5}$ (13 µg m^{-3})	-999	12.1
	1	14.8
	2	1.1
	3	16.2
	4	8.1
	5	9.1

4.1.2 Discussion

The first thing to note is that with these criteria, there are no synoptic types where high values never happen, and there are no synoptic types where high values always happen. Thus the synoptic type by itself is neither necessary nor sufficient to determine if a given day will have elevated levels of O_3 or $PM_{2.5}$. However, the frequency of elevated levels of O_3 and $PM_{2.5}$ are clearly different between the different synoptic types, and this pattern is occasionally different between the Group 1 urban areas.

The second thing to note is that unclassified days in the O_3 season (those that do not match any of the 5 synoptic types, and thus indicated as MT -999) have relatively high total and background O_3 values for all four urban areas. This may point to a limitation in our current map-typing scheme, which uses data from the entire year to determine the five most-frequent types. Further work is thus required to redo the classification for just the O_3 season and to increase the number of synoptic types.

Below we discuss the results for each of the pollutant metrics in turn.

4.1.2.1 Total MDA8 O_3

In HGB (Figure 21 and Table 7), the synoptic types can be sorted into two groups for the percentage of high values of total MDA8 O_3: a low group (MT 1, MT 3, and MT4) and a high group (MT 2, MT 5, and unclassified or "MT -999"). Overall, 32.3 % of days in the high group have total MDA8 O_3 above 70 ppbv, compared to only 13.7% of the days in the low group. MT 2 and MT 5 both feature stagnant high-pressure systems over Southeast Texas, and so it is not surprising that these clear days with low winds are favorable for O_3 production. MT 1 and MT 3 have flow coming from the Gulf of Mexico over HGB, and MT 4 has a relatively fast southwesterly flow, both of which would tend to reduce the O_3 levels in HGB. The percentage of unclassified days with high levels of total MDA8 O_3 is also high (32.6%), and far more days in the O_3 season fall into this type than into MT 2 or MT 5, so further work may be needed to classify these days into additional synoptic types for further analysis.

In DFW (Figure 22 and Table 8), the synoptic types fall into three groups with high (MT 5 and MT -999, 35-40%), medium (MT 2 and MT 3, 21-25%), and low (MT 1 and MT 4, 12-13%) percentages of days with total MDA8 O_3 above 70 ppbv. MT 2 has relatively fewer high values in DFW than it did in HGB, likely because the northwesterly flow in MT 2 is stronger near DFW than near the Gulf of Mexico. In addition, while the flow from the Gulf in MT 3 tends to reduce O_3 for HGB, it has less of an effect on DFW, likely because in this map type DFW receives outflow from most of eastern Texas.

In SA (Figure 23 and Table 9), MT 2 and MT -999 have the highest percentage of high values. MT 5 has high levels about half as frequently as MT 2, mainly because MT 5 is less stagnant over SA than MT 2, while both types have similar stagnation over HGB. The relatively fast flow of MT 1 over SA leads to relatively few high O_3 days.

The results for ARR (Figure 24 and Table 10) are similar to HGB in that MT 1, MT 3, and MT4 have a low percentage of high O_3 days (1-4%) while MT 2, MT 5, and MT -999 have a high percentage (8-18%). This is consistent with the meteorology affecting ARR being similar to that affecting HGB for these synoptic types.

4.1.2.2 Background MDA8 O_3

The background O_3 results for HGB follow a similar pattern to the total O_3 results discussed in Section 4.1.2.1, with MT 2, MT 5, and MT -999 having a relatively high percentage of high background O_3 days. This same pattern holds for DFW, SA, and ARR, and likely reflects the higher stagnation during these synoptic types. The fact that the synoptic types affect background O_3 similarly in all four urban areas suggests that the TCEQ method for calculating background O_3 (see Section 3.1 and Appendix B) is capturing regional, synoptic influences on O_3 in these urban areas.

4.1.2.3 Total Daily Average $PM_{2.5}$

In HGB (Figure 21 and Table 7), all of the synoptic types have similar percentages of days with total $PM_{2.5}$ above 17 ug/m^3 except for MT 2, which is significantly lower. As noted above, this is a relatively stagnant type over HGB with a slow northwesterly flow. SA (Figure 23 and Table 9) and ARR (Figure 24 and Table 10) are similar, with MT 2 as a clear outlier with a low percentage of high $PM_{2.5}$ days. The fact that the synoptic types with flow from the Gulf (MT 1 and MT 3) do not have appreciably lower $PM_{2.5}$ may be due to the transport of marine aerosol into the urban regions, as also suggested by our GAM fits (see Section A.1.5.2). Similarly, the fact that the $PM_{2.5}$ distributions are less sensitive to synoptic type than O_3 is consistent with our GAM results (Section 2.2.2), which showed that less of the variation of $PM_{2.5}$ could be attributed to meteorology than was the case for O_3.

In contrast, in DFW (Figure 22 and Table 8), MT 2, MT 4, and MT 5 all have relatively low frequencies of high $PM_{2.5}$ values. Both MT 4 and MT 5 have flow coming into DFW from the southwest, which gives les frequent high $PM_{2.5}$ values than the southerly to southeasterly flow in types MT 1 and MT 3.

4.1.2.4 Background Daily Average $PM_{2.5}$

In HGB (Figure 21 and Table 7), there is more difference between the types in terms of the percentage of days with high background $PM_{2.5}$ than there is for total $PM_{2.5}$. While MT 2 is again clearly lower than the other types, MT 1, MT 4, and MT 5 are in a group between MT 2 and the high group of MT 3 and MT -999. This again suggests that marine aerosol are a significant part of the background and total $PM_{2.5}$ in HGB when the flow is from the Gulf, as that flow dominates in MT 3 while the flow in MT 4 and MT 5 is from the southwest.

In DFW (Figure 22 and Table 8), there appear to be four groups: MT 2 at the low end (1%), MT 4 and MT 5 (6-9%), MT 1 and MT -999 (15-17%), and MT 3 (26%). This is again consistent with flow from the southwest bringing relatively lower background $PM_{2.5}$ to DFW than flow from eastern Texas. Austin/Round Rock (Figure 24 and Table 10) shows a similar dependence of the background $PM_{2.5}$ on the synoptic types to DFW.

SA (Figure 23 and Table 9) only has MT 2 as a clear low outlier, similar to the results for total $PM_{2.5}$.

4.2 Urban-Scale Meteorological Predictors of O_3 and $PM_{2.5}$

4.2.1 Logistic Regression Approach

One goal of this project (Deliverable 4.2) is to determine if there are necessary and/or sufficient synoptic or urban-scale meteorological criteria for events of "high" total and background O_3 and $PM_{2.5}$ (here we again define "high" using the criteria in Section 4.1.1). There

are likely no conditions where the probability of high O_3 and $PM_{2.5}$ is negligibly close to zero or one. Thus, in order to make our investigation of "necessary and/or sufficient" conditions for high O_3 and $PM_{2.5}$ tractable, we adopt the following probability definitions, recognizing that they are arbitrary choices:

- "Necessary" will refer to conditions that *must* be true for the probability of high O_3 and $PM_{2.5}$ (as defined in Section 4.1.1) to be greater than 20%.
- "More likely than not" will refer to conditions that, when true, give a greater than 50% chance of high O_3 and $PM_{2.5}$.
- "Sufficient" will refer to conditions that, when true, give a greater than 80% chance of high O_3 and $PM_{2.5}$.

Two ways to determine necessary and/or sufficient meteorological conditions have already been presented. First, the *gam03* and *back_gam03* models described in Sections 2.2.2 and 2.3 can be used to predict the actual values of total and background O_3 and $PM_{2.5}$ given the set of urban-scale meteorological predictors listed in Table A.7 and Table A.8. These predicted values and their confidence intervals can be used to estimate the probability that there will be high O_3 and $PM_{2.5}$ given a set of meteorological conditions. Second, in Section 4.1.2 we have shown that the probability of high O_3 and $PM_{2.5}$ events does vary between synoptic types.

Here we use the technique of logistic regression to create GAM models relating smooth functions of urban-scale and synoptic-scale meteorological variables to the probability that a high O_3 or $PM_{2.5}$ event will occur. Similar to the GAM equation in Section 2.2, the logistic regression equation is given by

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

where μ_i is the *i*th day's observation of whether or not a high O_3 or $PM_{2.5}$ event occurred (coded as 1 for true and 0 for false), $g(\mu_i)$ is the "link" function (here, a logit link is used with a binomial probability distribution, unlike the log link and Gaussian distribution used for the GAMs of Section 2.2), and $x_{i,j}$ are the n urban scale 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. We do not include the day of week, year, and day of year variables in our logistic regression. Instead, we include a factor (S_m) describing the synoptic types described in Section 4.1. 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).

In order to simplify our analysis, we focused on just two urban-scale meteorological predictors, afternoon mean temperature and daily average wind speed. These variables were chosen as they seemed to have the biggest impact on the predicted values of both total and background O_3 and $PM_{2.5}$ in our GAM fits from Sections 2.2.2 and 2.3. We then plot the probability of a high O_3 or $PM_{2.5}$ event estimated by the logistic regression equation as a function of afternoon mean temperature and daily average wind speed, with a separate plot for each combination of pollutant metric, Group 1 urban area, and synoptic type. This results in 16 figures with six panels in each figure. The plots of HGB are included as Figure 25 through Figure 28 in this section, while the plots for the other three Group 1 urban areas are included in Appendix D: Logistic Regression Probability Plots for DFW, SA, and ARR below.

One thing to note is that these logistic models and the associated figures can be used to forecast the probability of a high O_3 or $PM_{2.5}$ events based on a corresponding meteorological forecast. The forecast for geopotential height at 850 mbar can be used to determine the synoptic

type, and then the forecast of 10 m winds and 2 m temperatures at the locations listed in Table A.1 can be used to estimate the probability of an event. Further work should explore this approach further to determine the performance of such forecasts.

Table 11. Deviance explained (%, bold) and URBE score (unitless, italics) for the logistic models for total and background O_3 and $PM_{2.5}$

Urban Area	Total MDA8 O_3	BG MDA8 O_3	Total Daily $PM_{2.5}$	BG Daily $PM_{2.5}$
HGB	**25.7** *-0.187*	**14.4** *-0.450*	**9.7** *-0.035*	**10.4** *-0.239*
DFW	**35.0** *-0.243*	**18.3** *-0.186*	**14.7** *-0.341*	**18.0** *-0.363*
SA	**22.4** *-0.480*	**13.7** *-0.310*	**15.0** *-0.432*	**12.8** *-0.175*
ARR	**17.3** *-0.565*	**14.2** *-0.193*	**13.3** *-0.487*	**13.8** *-0.397*

4.2.2 Results and Discussion

The percent of the deviance[2] explained by the logistic model and the Un-Biased Risk Estimator[3] (UBRE) score for each logistic model is given in Table 11. The models explain the largest percentage of deviance for total MDA8 O_3, but even here only 17-35% of the deviance is explained, suggesting most of the variability is due to other parameters not included in the model. Additional meteorological predictors from Table A.7 and Table A.8 could be added to increase the amount of deviance explained, but adding predictors would make interpreting the results in terms of necessary and sufficient conditions more difficult.

The daily average wind speed is always a significant predictor at the $\alpha = 0.001$ level, and the afternoon temperature is always significant at the $\alpha = 0.005$ and significant at the $\alpha = 0.001$ level except for the background O_3 fits for HGB and SA. The differences between the factors for the synoptic types are occasionally significant, but many types are found to be similar to each other, as expected from our discussion in Section 4.1.2.

Below we discuss the results for each of the pollutant metrics in turn.

4.2.2.1 Total MDA8 O_3

As expected, the general trend for total MDA8 O_3 in all four Group 1 urban areas is to increase with increasing afternoon mean temperature and decreasing with daily average wind speed. We focus first on the HGB (Figure 25) and DFW (Figure D.1) urban areas, as these are the only urban areas that have meteorological conditions where the probability of total MDA8 O_3 being above 70 ppbv is greater than 80%, and is thus "sufficient" under our definition. However, we must also note that these are the best-monitored urban areas as well, which may influence the high probabilities.

[2] "Deviance" plays a similar role in GAMs as the variance of the residuals in ordinary linear models (see Wood, 2006, p. 70 for the full definition). The percent of deviance explained by a GAM is a generalization of r^2 for ordinary linear models (Wood, 2006, p. 84).

[3] For logistic regression with GAMs, minimizing the UBRE score (see Wood, 2006, p. 172 for the full definition) is equivalent to minimizing the expected mean square error of the model. The lower the score, the better the model fit.

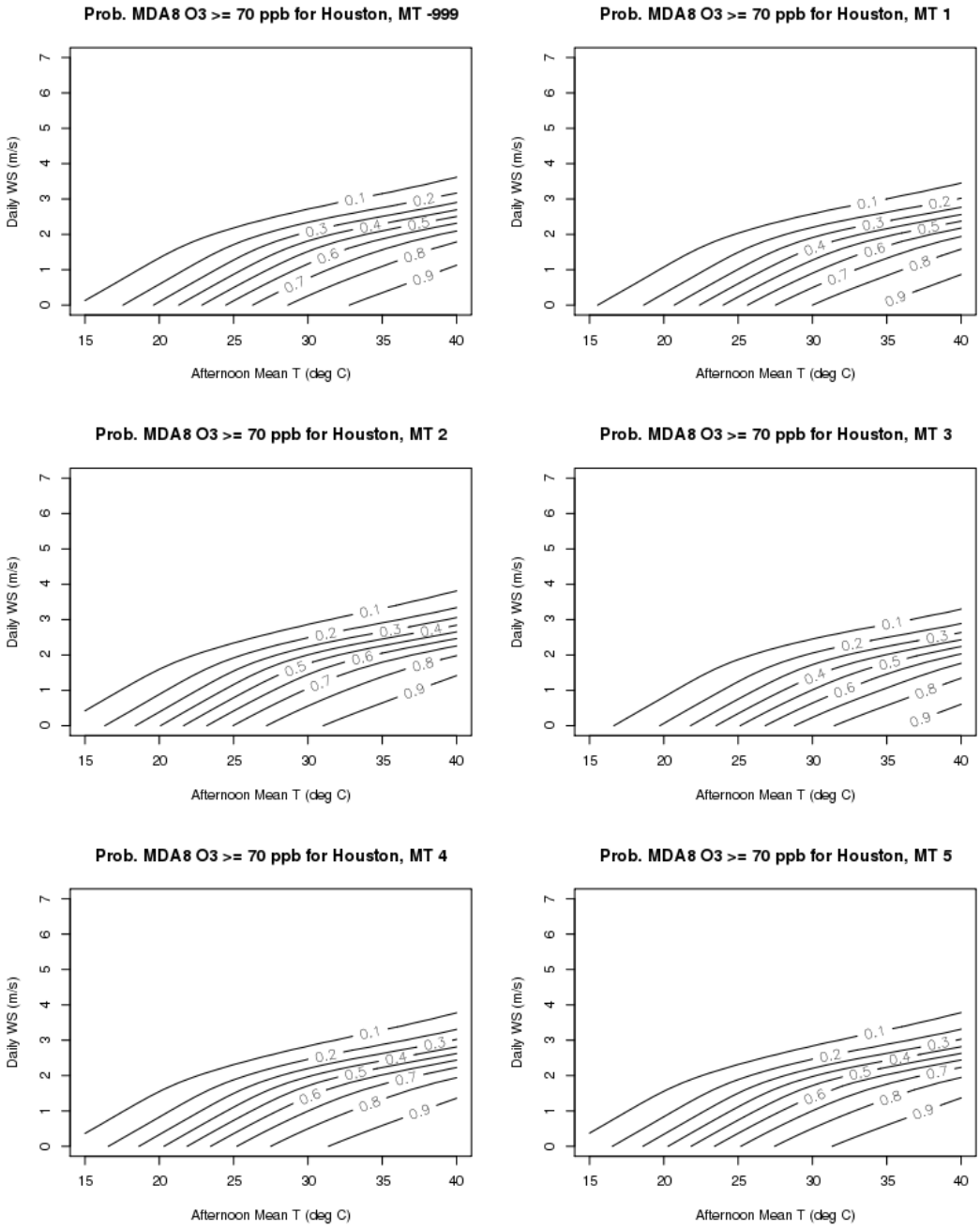


Figure 25. Probability of the total MDA8 O_3 exceeding 70 ppbv for the Houston/Galveston/Brazoria urban area as a function of afternoon mean temperature (°C), daily wind speed (m/s), and synoptic type (as defined in Section 4.1).

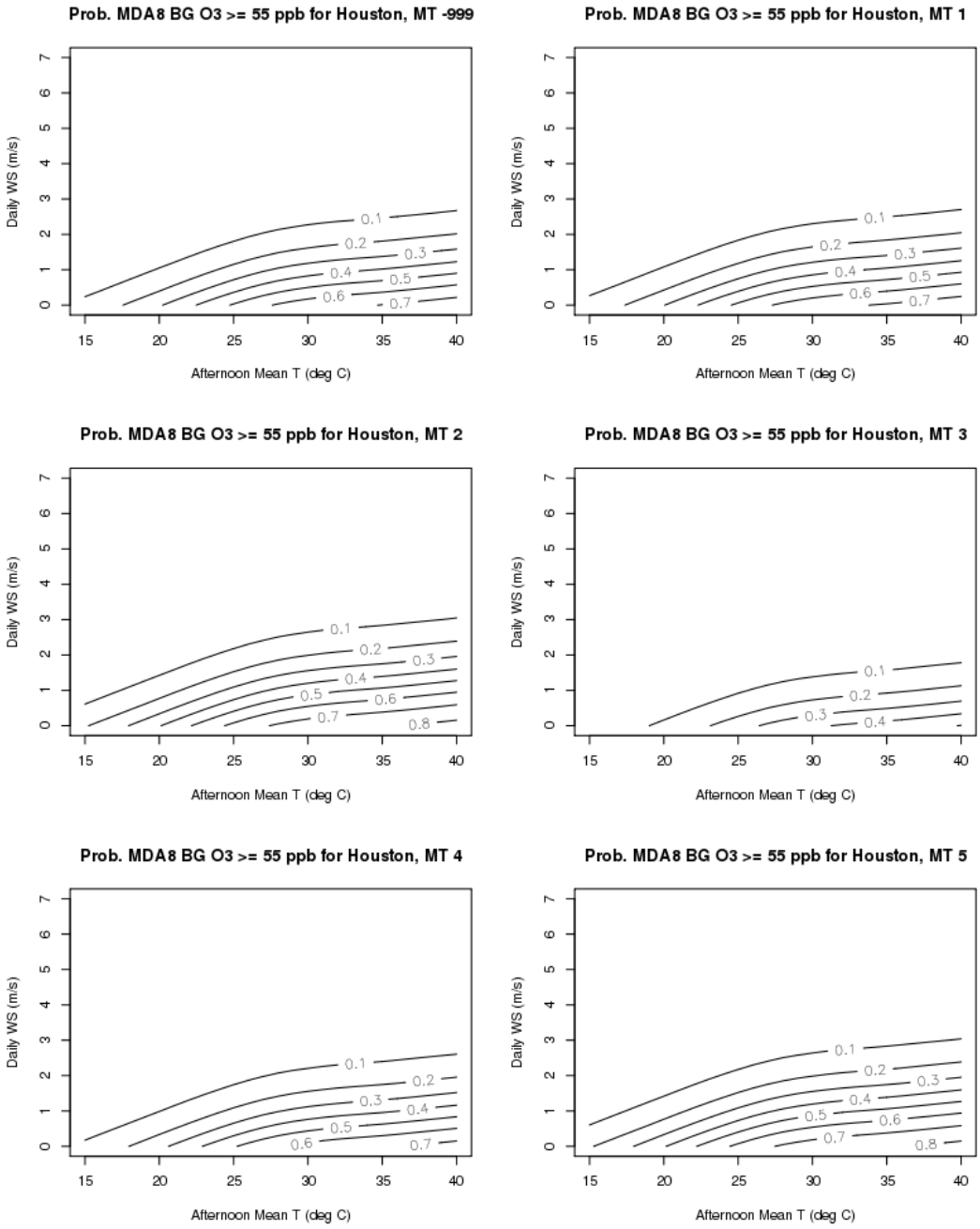


Figure 26. As in Figure 25, but for the probability that background MDA8 O_3 will exceed 55 ppbv.

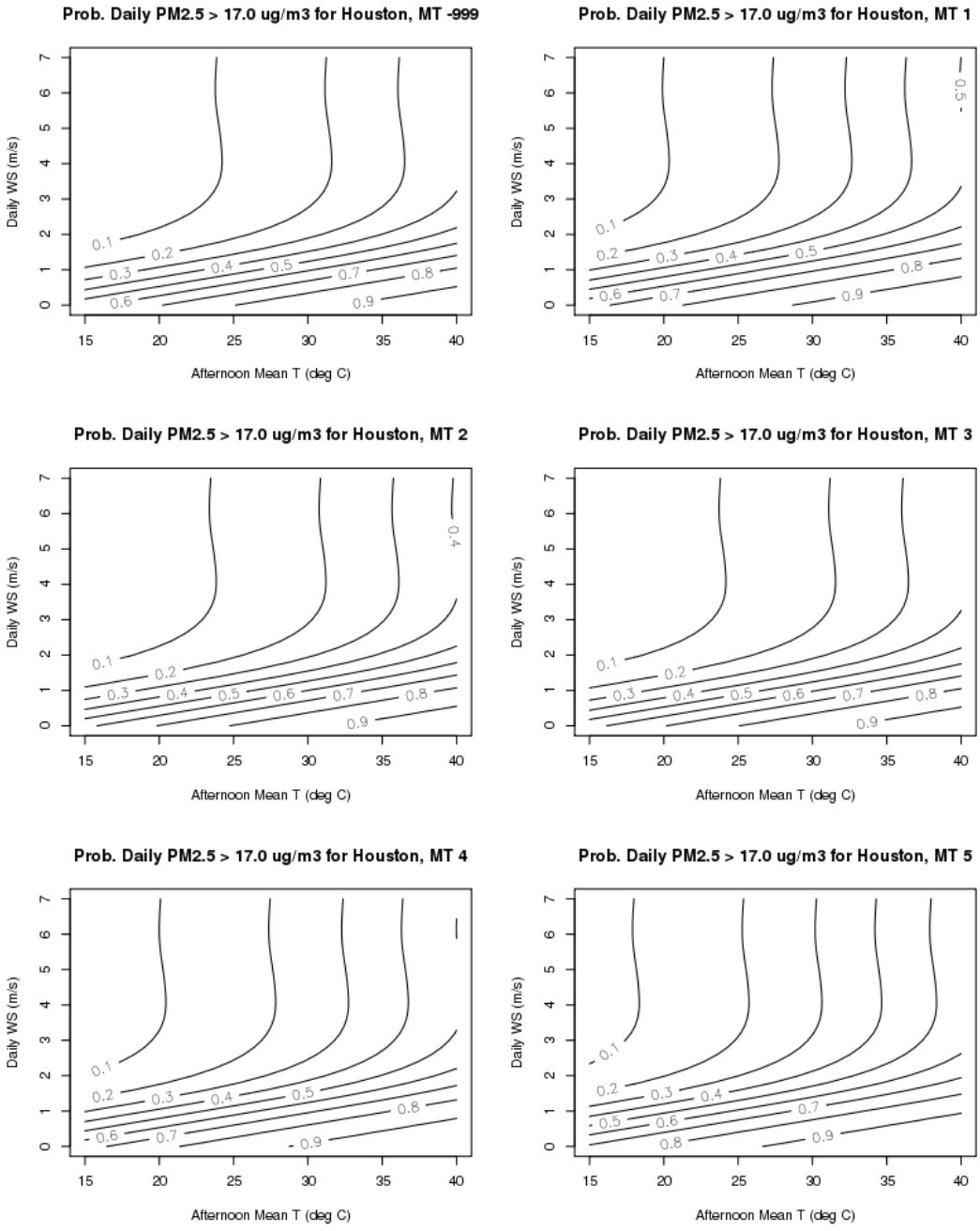


Figure 27. As in Figure 25, but for the probability that total daily average $PM_{2.5}$ will exceed 17 $\mu g/m^3$.

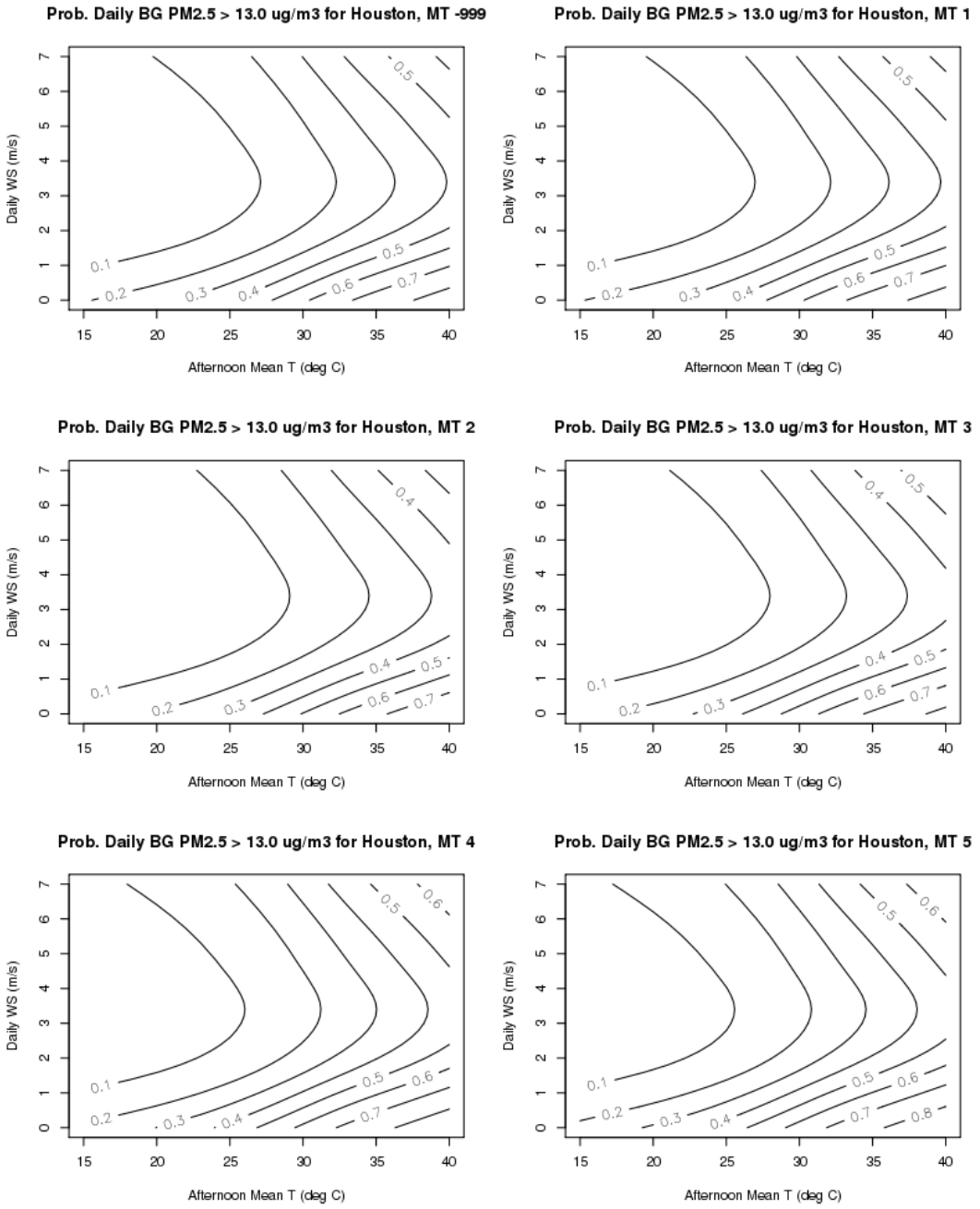


Figure 28. As in Figure 25, but for the probability that background daily average $PM_{2.5}$ will exceed 13 $\mu g/m^3$.

For HGB and DFW, once you correct for the relative distribution of mean afternoon temperature and average wind speed in each synoptic type, the synoptic types only account for a minor change in the probabilities. This suggests that the patterns seen for MDA8 O_3 with synoptic type in Section 4.1.2.1 are mainly due to differences in the temperature and wind speed distributions within each type.

In HGB, the "necessary" conditions (probability greater than 20%) for high total O_3 are afternoon temperatures above 17 °C and average wind speeds below 3.5 m/s. In contrast, DFW requires temperatures above 20 °C for the odds of a high total O_3 event to get above 20%, but the wind speed can be as high as 4 m/s. The "sufficient" conditions (probability greater than 80%) in HGB are afternoon temperatures above approximately 29 °C and average wind speeds less than approximately 1.5 m/s. For DFW, the conditions are also temperatures above approximately 29 °C and the wind speed below 1-2 m/s, with the exact value depending on synoptic type. The gradients of the probabilities with respect to afternoon temperature and average wind speed look very similar between HGB and DFW.

For SA (Figure D.5) and ARR (Figure D.9), we should first note that the minimum and maximum measured average wind speeds in these areas were approximately 3.5 and 6 m/s, respectively, so data at these higher wind speeds is an extrapolation and should not be trusted.

In SA, the probabilities of high total O_3 events are a much stronger function of wind speed than of temperature, unlike in HGB and DFW where the two predictors had more equal influence. The "necessary" conditions are average wind speed below approximately 1.5 m/s and afternoon temperatures above approximately 21 °C. The "more likely than not" conditions (probability greater than 50%) are that you must be in synoptic types MT -999, MT 2, MT 4, or MT 5 (i.e., the synoptic flow is not from the Gulf of Mexico as in MT 1 and MT 3), average wind speeds below approximately 0.5 m/s, and afternoon temperatures above 25-29 °C, with the critical temperature varying with synoptic type.

In ARR, the probability of a high total O_3 event never gets above 20% for MT 1 (southerly flow from the Gulf) and never gets above 50% for MT 3 (southwesterly flow from the Gulf). The "necessary" conditions are thus a synoptic type other than MT 1, afternoon temperatures above 24-27 °C and wind speeds below 1-2 m/s, depending on synoptic type. The "more likely than not" conditions are near zero wind speeds and afternoon temperatures above 33 °C for MT 4 (fast southeasterly flow) and wind speed below 1 m/s and afternoon temperatures above 27 °C for MT -999, MT 2, or MT 5.

4.2.2.2 Background MDA8 O_3

As expected, the patterns in the probabilities of high background O_3 are similar to the patterns for total O_3 for each urban area. For HGB (Figure 26), the "necessary" conditions are afternoon temperatures above approximately 17 °C and average wind speeds below approximately 2.5 m/s for all synoptic types except MT 3 (southeasterly Gulf flow). For MT 3, the temperature must be above 23 °C and the wind speeds below 1 m/s. A high background O_3 event is "more likely than not" for temperatures above 23-25 °C and wind speeds below 1 m/s for all synoptic types except MT 3, where calm conditions and temperatures above 32 °C are required. "Sufficient" conditions for high background O_3 are rare, only occurring for temperatures above 36 °C for MT 2 and MT 5.

For DFW, the "necessary" conditions are wind speeds below approximately 3 m/s and afternoon temperatures above 18 °C (MT 2 only) or above 20 °C (all other types). The "sufficient" conditions are wind speeds near 0 m/s and afternoon temperatures above 28 °C.

For SA, probabilities of high background O_3 events over 50% are extremely rare, and the probabilities are mainly a function of wind speed and synoptic type. Wind speeds below 1 m/s are "necessary" for an event to occur.

For ARR, the apparent second maxima at high wind speeds is likely an artifact of the model extrapolating to wind speeds not included in the actual observations. The probabilities for a high background O_3 event are a strong function of synoptic type in this urban area, and are highest for MT -999 and 2 (unclassified and stagnant conditions) and lowest for MT 1 and 3 (flow from the Gulf).

- For MT 1 and MT 3, the "necessary" conditions are temperatures above 25 °C and wind speeds below 2 m/s, with no conditions reaching 50% probability.
- For MT 4 and MT 5 (southeasterly flow), the "necessary" conditions are temperatures above 22 °C and wind speeds below 3 m/s, while a high event is "more likely than not" for wind speeds below 1.5 m/s and temperatures above 27 °C.
- For MT -999 and MT 2 (unclassified and stagnant conditions), the "necessary conditions are temperatures above 19-21 °C with a weak dependence on wind speed. A high background O_3 event is "more likely than not" for temperatures above 23-25 °C and wind speeds below 2 m/s.

4.2.2.3 Total Daily Average $PM_{2.5}$

The dependence of the probability of a high total $PM_{2.5}$ event on daily average wind speed is dramatically different than that for high total O_3 events for all four urban areas. HGB and DFW (Figure 27 and Figure D.3) both show the probabilities increasing with decreasing wind speed for wind speeds below 2-3 m/s, but the probabilities become independent of wind speed at higher wind speeds. In SA (Figure D.7), the probabilities mainly depend on wind speed as was the case for O_3, but the plots show a secondary maxima for afternoon temperatures near 34 °C and average wind speeds near 2.5 m/s. In ARR (Figure D.11), the probabilities are rarely above 20% and the patterns are likely due to numerical errors, and so are not discussed any further.

For HGB, when the wind speeds are above approximately 3 m/s, the "necessary" conditions are afternoon temperatures above 30 °C for MT -999, MT 2, and MT 3, above 27 °C for MT 1 and MT4, and above 25 °C for MT5. For wind speeds below 3 m/s, the necessary conditions are wind speeds below 2 m/s. The "sufficient" conditions are temperatures above 20-22 °C (MT 5 and MT1), 25 °C (MT -999, 2, and 3) or 29 °C (MT 4) and wind speeds below 1 m/s.

For DFW, when the wind speeds are above approximately 3 m/s, the "necessary" conditions are afternoon temperatures above 30 °C for MT 1, above 32 °C for MT 3, and above 35 °C (MT -999, 54, 5). There are no "sufficient" conditions for a high $PM_{2.5}$ event for MT 2 (stagnant conditions). For MT 1, temperatures above 30 °C and wind speeds below 1 m/s are "sufficient", and temperatures above 35 °C with wind speeds near 0 m/s are "sufficient" for the other synoptic types.

There are no "sufficient" conditions for a high $PM_{2.5}$ event in SA. For MT 2, the "necessary" conditions are temperatures above 26 °C and wind speeds near 0 m/s. For the other synoptic types, the "necessary" conditions are wind speeds below 1 m/s, with a high $PM_{2.5}$ event being "more likely than not" for temperatures above 27 °C and wind speeds near 0 m/s. In addition, the

"necessary" conditions for a high $PM_{2.5}$ event exist for wind speeds around 2 m/s and temperatures around 35 °C, with the size of this secondary maximum depending on the synoptic type.

4.2.2.4 Background Daily Average $PM_{2.5}$

For HGB (Figure 28), we see a parabolic dependence of the probability of a high background $PM_{2.5}$ on the daily average wind speed with a minimum around 3.5 m/s. The increased probability of a high background $PM_{2.5}$ event at high wind speeds likely reflects an increase in marine or dust aerosol emission at higher wind speeds. The "necessary" conditions for an event are afternoon temperatures above 15-19 °C during calm conditions, temperatures above 30-34 °C at the wind speed minimum at 3.5 m/s, and temperatures above 25-27 °C at high wind speeds (greater than 6 m/s). There are few "sufficient" conditions for an event, but a high background $PM_{2.5}$ event is "more likely than not" for temperatures above 27-29 °C for calm conditions and above 35 °C for fast wind speeds.

For DFW (Figure D.4), the probability of an event is below 20% or MT 2. For the other types, when wind speeds are greater than 2 m/s, the probability of a high $PM_{2.5}$ event is a function of afternoon temperature and synoptic type only, with the threshold temperatures at 27 °C (MT 1), 31 °C (MT -999 and MT 3), 33 °C (MT 5) and 35 °C (MT 4). At wind speeds below 2 m/s, the probabilities are strong functions of both temperature and wind speed. No conditions are "sufficient", but an event is "more likely than not" for temperatures above 28-33 °C and wind speeds below 1-2 m/s, with the critical values depending on the synoptic type.

For SA (Figure D.8), the data for temperatures above 35 °C are extrapolations as the temperature only rarely gets this high. As in total $PM_{2.5}$, we see a secondary probability maximum around 3 m/s and 35 °C. "More likely than not" conditions rarely occur, and the "necessary" conditions are largely independent of wind speed. There are no "necessary" conditions for MT 2, and the critical temperature for the other types varies between 26-29 °C.

As for total $PM_{2.5}$, the probabilities for a high background $PM_{2.5}$ event in ARR are rarely above 20%, and so are not discussed further.

5 Quality Assurance Steps and Reconciliation with User Requirements

All work on the project was done in accordance with the Quality Assurance Project Plan (QAPP). All scripts and data files used in this project were inspected by team members different from the original author to ensure they were correct, and any errors noted in early versions were fixed. Other required evaluations are contained within the report (for example, see Section A.1.6) 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.

In addition, the QAPP listed several questions that needed to be addressed for each project task. These questions are addressed below.

5.1 Task 2: Development of GAMs

- *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 A.1.4.2 and A.1.5.2, 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 A.1.4.2, 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 °C 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 A.1.2, 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 approximately 100 km and differences in bearing of less than approximately 20° are unlikely to be significant.

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

As noted in Section A.1.6, the two-fold cross-validation showed that the GAMs fit to 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 A.1.6, 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 identified any set of necessary or sufficient conditions that lead to large residuals in the GAMs.

5.2 Task 3: Background O_3 and $PM_{2.5}$

- *Are the derived background estimates, and their spatial and temporal variation, consistent with our conceptual models of O_3 and $PM_{2.5}$ emissions, chemistry, and transport?*

  The overall trends of background O_3 and $PM_{2.5}$ are decreasing, consistent with our understanding of reduction of pollutant emissions (primarily NO_x and SO_2) over this time period. Background O_3 has a minimum in July and a maximum in September for urban areas near the Gulf of Mexico, consistent with the seasonal shifts in synoptic conditions.

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

  The main literature comparison is the study of HGB background O_3 from 2000-2012 by Berlin et al. (2013). They found that unadjusted total MDA8 O_3 in HGB decreased at a rate of -0.89±0.66 ppbv/year from 2000 to 2012. Our calculated annual averages for the MDA8 O_3 for all HGB sites from 2005-2014 (see Figure 1 and Figure 8) show the same year-to-year variation as seen in that study during the overlapping period, but our average values are lower, probably due to different sets of sites (and those sites' continuous or non-continuous records) being included in the two studies. In this study, we chose sites that had only a continuous dataset during the ozone season from 2005-2014 for a total of 25 sites in Houston, whereas Berlin et al. (2013) had two sets of sites for their evaluations; a 6- and 30-site analysis, where their 6-site analysis had a continuous data set, and the 30-site analysis had a combination of continuous and non-continuous datasets. Our unadjusted trend for 2005-2014 (-1.48±0.73 ppbv/year) is consistent with their estimate, especially as 2013 and 2014 had lower average MDA8 O_3 than any of the years from 2005-2012.

  Similarly, the year to year variability of our TCEQ background O_3 estimates for HGB are consistent with those from Berlin et al. (2013), with no evidence of an

"offset" as seen in the total MDA8 O_3 values. Our trend estimate for 2005-2014 (-0.91±0.57 ppb/year) and the Berlin et al. (2013) estimate for 2000-2012 (-0.21±0.39 ppbv/year) are consistent to within the 95% confidence intervals, especially considering the low values seen in 2013 and 2014.

The trends in our PCA-based background O_3 estimates (see Figure 13) are similar to the 6-station PCA results from Berlin et al. (2013) in that both show a slight decreasing trend. However, differences exist in both the year-to-year variations and the overall magnitude of this background, likely due to our consideration of 25 sites for the entire May-October O_3 season. Additionally, we had separated the O_3 season, due to a mid-season shift in the Gulf prevailing winds. As noted below, further research is needed to understand the differences between the TCEQ and PCA estimates of background O_3.

- *What are the uncertainties in the background estimates, and under what conditions are they valid?*

  The major uncertainties in the background estimates calculated using the TCEQ method are, first, that they assume the regional background can be estimated as the lowest value observed at a selected number of sites around the urban area. This neglects the fact that the urban areas in Texas likely influence each other's "background", and so our background estimates cannot be interpreted as estimates of what the concentrations would be with all Texas sources removed. A second uncertainty is that the different urban areas have very different numbers of monitoring sites, and so the regional background is likely under-sampled for some urban areas, especially the Group 2 areas. So long as these caveats are kept in mind, the values should be valid throughout 2005-2014, except for $PM_{2.5}$ in the TLM urban area, as noted in Section B.2.4.

  The PCA-derived background estimates from Section 3.3 should be considered as experimental, but the estimates for O_3 show good correlation with the TCEQ-based estimates. Further research is needed to understand the differences between the two different estimates of background O_3.

- *Are the derived background estimates, and their spatial and temporal variation, consistent with the other data-based methods explored in this task? If not, is there a reasonable explanation for the differences?*

  As noted in Section 3.3.1, the TCEQ and PCA-based estimates of background O_3 are well-correlated with each other (r^2 greater than 0.79), and the year-to-year variability in the background O_3 distributions also appear consistent between the two methods. However, the slopes of the linear relationships between the two estimates differ between urban areas for reasons that are still unclear.

5.3 Task 4: Synoptic and Mesoscale Controls of O_3 and $PM_{2.5}$

- *Are the derived synoptic types and identified mesoscale meteorological controls on extreme and background concentrations of O_3 and $PM_{2.5}$ consistent with our conceptual understanding of O_3 and $PM_{2.5}$ emissions, chemistry, and transport?*

  The synoptic types identified in Section 4.1 appear to be reasonable, including a mixture of stagnant conditions, flow from the Gulf of Mexico, flow from the Western US, and flow from the southeast. The dependence of O_3 and $PM_{2.5}$ on

urban-scale meteorological predictors and on synoptic types is consistent with our understanding of O_3 and $PM_{2.5}$, as discussed in Sections 2.2, 4.1.2, 4.2.2, and A.1.5.2.

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

  The GAM results are consistent with Camalier et al. (2007) as noted above, and the synoptic type analysis is similar to that performed in Hegarty et al. (2007). The results of the logistic regressions of Section 4.2.2 are reasonable, except for some extrapolated conditions noted in the text.

6 Conclusions

Here we summarize the conclusions of our analysis, with reference to the corresponding report section and project deliverable.

- The Generalized Additive Models (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 GAMs also generally show good fits with normally-distributed residuals and little dependence of the residual variance on the predicted value (Sections 2.2.2.1 and A.1.5.2, Deliverable 2.2).
- 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 (Sections 2.2.2.2 and A.1.5.2, Deliverable 2.2).
- 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 $PM_{2.5}$ (Sections 2.2.2.1, 2.2.2.2, and A.1.5.2, Deliverable 2.2).
- 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 °C 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 speeds above 5 m/s and a minimum in $PM_{2.5}$ at a HYSPLIT bearing of 120° instead of at 320° (Sections 2.2.2.1, 2.2.2.2, and A.1.5.2, Deliverable 2.2).
- We calculated estimates of total and background MDA8 O_3 and daily average $PM_{2.5}$ for the period 2005-2015, with the background estimated calculated using the TCEQ method described in Berlin et al. (2013) (Sections 3.1 and B.2, Deliverable 3.1).
- We find that the meteorological relationships determined by fitting GAMs to background O_3 and $PM_{2.5}$ are substantially identical to those fit to total O_3 and $PM_{2.5}$, with the possible exception of HGB O_3 (Section 2.3, Deliverables 3.2 and 4.2).
- After meteorological adjustment via the GAMs fit to total and background O_3 and $PM_{2.5}$ for each urban area, several negative trends in pollutant metrics between 2005-2014 were observed to be significant at a 95% confidence level and no positive trends were observed (Section 2.4, Deliverables 2.2, 3.2, and 4.2).
    - For HGB, total and background MDA8 O_3 decreased at -0.46±0.40 ppbv/year and -0.41±0.31 ppbv/year, respectively, while total and background daily average $PM_{2.5}$ decreased by -0.39±0.14 μg/m^3/year and -0.37±0.11 μg/m^3/year, respectively.
    - For DFW, total and background MDA8 O_3 decreased at -0.68±0.39 ppbv/year and -0.57±0.49 ppbv/year, respectively, while total and background daily

average $PM_{2.5}$ decreased by -0.15±0.09 μg/m^3/year and -0.23±0.08 μg/m^3/year, respectively.

- In SA, background O_3 and $PM_{2.5}$ decreased by -1.01±0.87 ppbv/year and -0.12 ±0.07 μg/m^3/year, respectively.
- In ARR, background $PM_{2.5}$ decreased by -0.21±0.07 μg/m^3/year.
- In BPA, total $PM_{2.5}$ decreased by -0.34±0.11 μg/m^3/year.
- In TLM, total O_3 and $PM_{2.5}$ decreased by -0.66±0.44 ppbv/year and -0.34 ±0.08 μg/m^3/year, respectively.

- Background MDA8 O_3 is fairly constant with month during the O_3 season for TLM and DFW, but has a July minimum for the other urban areas. In contrast, median background $PM_{2.5}$ peaks in June and July. The range of values for a given month or year is large for all cities, with HGB having the largest O_3 spread and the most outliers, possibly due to the fact that it has the largest dataset (Section 3.2, Deliverable 3.2)
- We find that the principal component analysis (PCA) based method of Langford et al. (2009) can give reasonable values for background O_3 in the four Group 1 urban areas as long as the PCA analysis is performed separately for the May-July and August-October halves of the period. These PCA-based background estimates are well-correlated with the values derived using the TCEQ method, and show similar seasonal and inter-annual variability. However, the slope of the linear relationship between the two background O_3 estimates varies substantially between urban areas for reasons that are currently unclear and require further analysis (Section 3.3.1, Deliverable 3.3).
- In contrast, the PCA-based background estimates for $PM_{2.5}$ give unphysical values regardless of what time period is used for the analysis (Section 3.3.1.2, Deliverable 3.3).
- Currently existing methods, such as those of van Donkelaar et al. (2013), exist for using satellite observations to determine surface $PM_{2.5}$ concentrations at a high (0.01^o latitude by 0.01^o longitude) horizontal resolution. These methods should be further explored as an additional way of estimating regional background $PM_{2.5}$ (Section 3.3.4, Deliverable 3.3).
- In contrast, current satellite techniques have difficulty separating the boundary-layer O_3 mixing ratios from the free tropospheric values, but techniques that combining different satellite instruments (e.g., Fu et al., 2013) or use future instruments may make this possible (Section 3.3.3, Deliverable 3.3).
- We identified 5 synoptic types based on the NARR 850 mbar geopotential height fields that allow for the classification of 70% of all day in the 10-year study period and 58% of the days in the May-October O_3 season. However, the remaining unclassified days account for a significant proportion of high O_3 and $PM_{2.5}$ events, which suggests that future work should adjust the method of Hegarty et al. (2007) should be adjusted to classify more days (Section 4.1.1, Deliverable 4.1).
- We defined criteria for "high" levels of total and background MDA8 O_3 and daily average $PM_{2.5}$ based on the observed 90th percentile values of the Group 1 urban

areas. These criteria are 70 ppbv for total MDA8 O_3, 55 ppbv for background MDA8 O_3, 17.0 µg m^{-3} for total daily average $PM_{2.5}$, and 13.0 µg m^{-3} for background daily average $PM_{2.5}$ (Section 4.1.1, Deliverable 4.1).

- We found that the relative percentage of "high" O_3 and $PM_{2.5}$ events varied between the synoptic types and urban areas (Section 4.1.2, Deliverable 4.2).
    - For HGB, the stagnant types MT 2 and MT 5, and the unclassified "MT -999" days, had high percentages (greater than 30%) of days with events of high total and background O_3, but type MT 2 had significantly fewer high $PM_{2.5}$ events (14% for total $PM_{2.5}$) than the other synoptic types (19-27%).
    - For DFW, types MT 5 and MT -999 have a very high percentage (greater than 35%) of days with high total O_3 and the percentage of background O_3 events is high for MT2, MT 5, and MT -999 (greater than 25%). MT 1 and MT 3,which have southerly and southeasterly flow, as well as the unclassified days in MT -999, have a relatively high percentage of high $PM_{2.5}$ events (15-25%).
    - For SA, types MT 2 and MT -999 are associated with high O_3 events (16-19% for total, 21-22% for background) and type MT 2 has an unusually low number of high $PM_{2.5}$ events (2-3% versus 7-25% for the other types).
    - ARR has similar O_3 results to HGB, with MT 2, MT 5, and MT -999 relatively high (8-18% for total, 20-33% for background), with MT 2 having a relatively low percentage of high $PM_{2.5}$ events (1% versus 6-13%).
- We performed logistic regressions to determine how the probability of high O_3 and $PM_{2.5}$ event in each urban area changed with afternoon mean temperature, daily average wind speed, and synoptic type. These predictors were chosen as they had been shown to be important in our GAM models and our previous analysis of the synoptic types. We used these probability models to investigate "necessary" (defined as giving a probability of a high event greater than 20%) and "sufficient" (defined as giving a probability of a high event greater than 80%) criteria for high O_3 and $PM_{2.5}$ events (Section 4.2, Deliverable 4.2).
    - For HGB, "sufficient" conditions for high total O_3 are afternoon temperatures above 29 °C and wind speed below 1-2 m/s depending on synoptic type. High $PM_{2.5}$ events can occur at both low and high wind speeds, but "sufficient" conditions include wind speeds below 1 m/s and afternoon temperatures above a critical value that varies between 20-29 °C with synoptic type.
    - For DFW, "sufficient" conditions for high total and background O_3 are wind speeds below 1-2 m/s and afternoon mean temperatures above approximately 29 °C, with the exact values depending on synoptic type. "Sufficient" conditions for high total $PM_{2.5}$ are temperatures above 30-35 °C depending on synoptic type.
    - For SA, there are no "sufficient" conditions for high O_3 or $PM_{2.5}$ events. High total O_3 events are "more likely than not" when the synoptic flow is not from the Gulf and the afternoon temperatures are above 25-29 °C. High total $PM_{2.5}$ events are "more likely than not" when temperatures are above 27 °C and wind speeds are near zero.

- For ARR, high total O_3 events are "more likely than not" for wind speeds near 0 m/s and afternoon temperatures above 33 °C for MT 4 (fast southeasterly flow) and wind speed below 1 m/s and afternoon temperatures above 27 °C for the relatively stagnant conditions of MT -999, MT 2, or MT 5. The probability of high $PM_{2.5}$ events is generally less than 20% under all conditions.

7 Recommendations for Future Study

As already noted in the report, there are several questions raised by the results of our current study that would benefit from further investigation. For example, not all of the urban areas in our study were equally well-sampled for either air quality or meteorological parameters. While this is unavoidable for a historical data study like this, future work could quantify the impact of the relative sparsity of observations for some urban areas on the robustness of our conclusions, especially those about differences between urban areas. For example, further work is needed to determine if the slope of the linear relationship between the TCEQ method and PCA-based method background O_3 estimates varies substantially between urban areas because of actual differences between the urban areas or differences in their monitoring networks.

In addition, the synoptic typing method used in this study has provided valuable information on the dependence of high O_3 and $PM_{2.5}$ events on synoptic conditions, but the fact that 30% of all days and 42% of days in the O_3 season are not covered by the five current types, and the fact that these unclassified days have a relatively high percentage of high O_3 events, suggests the need for further refinements to the synoptic typing technique to classify more of the remaining days.

We also discussed how satellite observations could be used to derive high-resolution estimates of surface $PM_{2.5}$ concentration, which could be used to refine regional background estimates. Future work should pursue this possibility and compare the derived background estimates with in situ measured values in Texas.

The GAMs developed in this study to relate meteorological predictors to the concentrations of total O_3 and $PM_{2.5}$, as well as the logistic GAMs used to determine necessary and sufficient conditions for high O_3 and $PM_{2.5}$ events, should be further developed and refined to provide accurate forecasts of air quality for the urban areas studied in this project.

Finally, these GAMs derived from monitor network data should be compared with similar GAMs fit to meteorological and chemical data from 3D Eulerian air quality models like CAMx and CMAQ to determine if these models accurately represent the dependence of O_3 and $PM_{2.5}$ concentrations, and the probability of high O_3 and $PM_{2.5}$ events, on meteorology. Differences discovered between the two sets of GAMs could point towards missing physics or incorrect parameterizations in the current Eulerian air quality models.

8 References

Berlin, S. R., A. O. Langford, M. Estes, M. Dong, and D. D. Parrish (2013) Magnitude, decadal changes, and impact of regional background ozone transported into the greater Houston, Texas, area, *Environ. Sci. & Technol.*, 47, 13985-92.

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.

Fu, D., Worden, J. R., Liu, X., Kulawik, S. S., Bowman, K. W., and Natraj, V.: Characterization of ozone profiles derived from Aura TES and OMI radiances, *Atmos. Chem. Phys., 13*, 3445-3462, doi:10.5194/acp-13-3445-2013, 2013.

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

Hegarty, J., H. Mao, *and* R. Talbot *(*2007*),* Synoptic controls on summertime surface ozone in the northeastern United States, J. Geophys. Res., 112, *D14306, doi:*10.1029/2006JD008170.

Langford, A. O., J. Brioude, O. R. Cooper, C. J. Senff, R. J. Alvarez II, R. M. Hardesty, B. J. Johnson, and S. J. Oltmans (2012), Stratospheric influence on surface ozone in the Los Angeles area during late spring and early summer of 2010, J. Geophys. Res., 117, D00V06, doi:10.1029/2011JD016766.

R Core Team (2015), R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org.

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

van Donkelaar, A., R. V. Martin, R. J. D. Spurr, R. Burnett (2015), Combining GEOS-Chem, satellite, and ground monitors for improved PM2.5 population exposure estimates, 7th Annual GEOS-Chem Meeting, Harvard University, May 4th-7th, 2015

van Donkelaar, A., R. V. Martin, R. J. D. Spurr, E. Drury, L. A. Remer, R. C. Levy, and J. Wang (2013), Optimal estimation for global ground-level fine particulate matter concentrations, J. Geophys. Res. Atmos., 118, 5621–5636, doi:10.1002/jgrd.50479.

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

Zoogman, P., D.J. Jacob, K. Chance, H.M. Worden, D.P. Edwards, L. Zhang, (2014) Improved monitoring of surface ozone by joint assimilation of geostationary satellite observations of ozone and CO, Atmos. Environ. 84, 254-261.

Appendix A. Effects of Meteorology on O_3 and $PM_{2.5}$ Trends

This appendix documents the files provided to TCEQ to complete Deliverable 2.2 of Work Order No. 582-15-54118-01. 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/anonymous/pub/malvarad/p1952_deliverable_2_2_R1_0.tar.gz`

Our major findings are (see Sections A.1.4, A.1.5, and A.1.6 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 A.1 of this memo briefly outlines the technical approach used to prepare the generalized additive models (GAMs) in the deliverable and Section A.2 describes the files in the deliverable. Section A.3 briefly outlines the quality assurance steps that have been performed.

A.1 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 A.1. Surface meteorological sites selected for GAM fitting.. 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 GAM. 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 A.1.1 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 A.1.2 describes the generation and evaluation of the HYSPLIT back trajectories. Section A.1.3 gives an overview of our GAM fitting procedure, followed by an overview of the GAM results for both the baseline (Section A.1.4) and extended (Section A.1.5) GAMs. Section A.1.6 then presents the results of the cross-validation analysis of the "gam03_extended" GAMs from Section A.1.5.

A.1.1 Input Data and Processing

A.1.1.1 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 A.1 - they were selected to maximize the amount of data available at each site.

Table A.1. 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 the Appendix to the original deliverable, we developed a python script (*calc_bkgrd_ozone.py*, see Section A.2.2) 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. The TCEQ monitor data, Integrated Global Radiosonde Archive data (IGRA, Section A.1.1.2) and the integrated surface hourly (ISH) database of the National Climatic Data Center (NCDC, Section A.1.1.3), along with the previously calculated MDA8 and $PM_{2.5}$ maximum and minimum concentrations and parameter from the HYSPLIT back trajectories (Section A.1.2), 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.

A.1.1.2 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 A.2 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 A.2.1.1 describes these files in further detail.

Table A.2. 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

A.1.1.3 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 A.3 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 A.2.1.2.

Table A.3. 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

A.1.1.4 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.

A.1.2 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 A.1 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 A.2.3.3). 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 A.1 shows an example ensemble back-trajectory calculation for August 25, 2013 in HGB, 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 approximately 100 km and differences in bearing of less than approximately 20^{o} are unlikely to be significant.

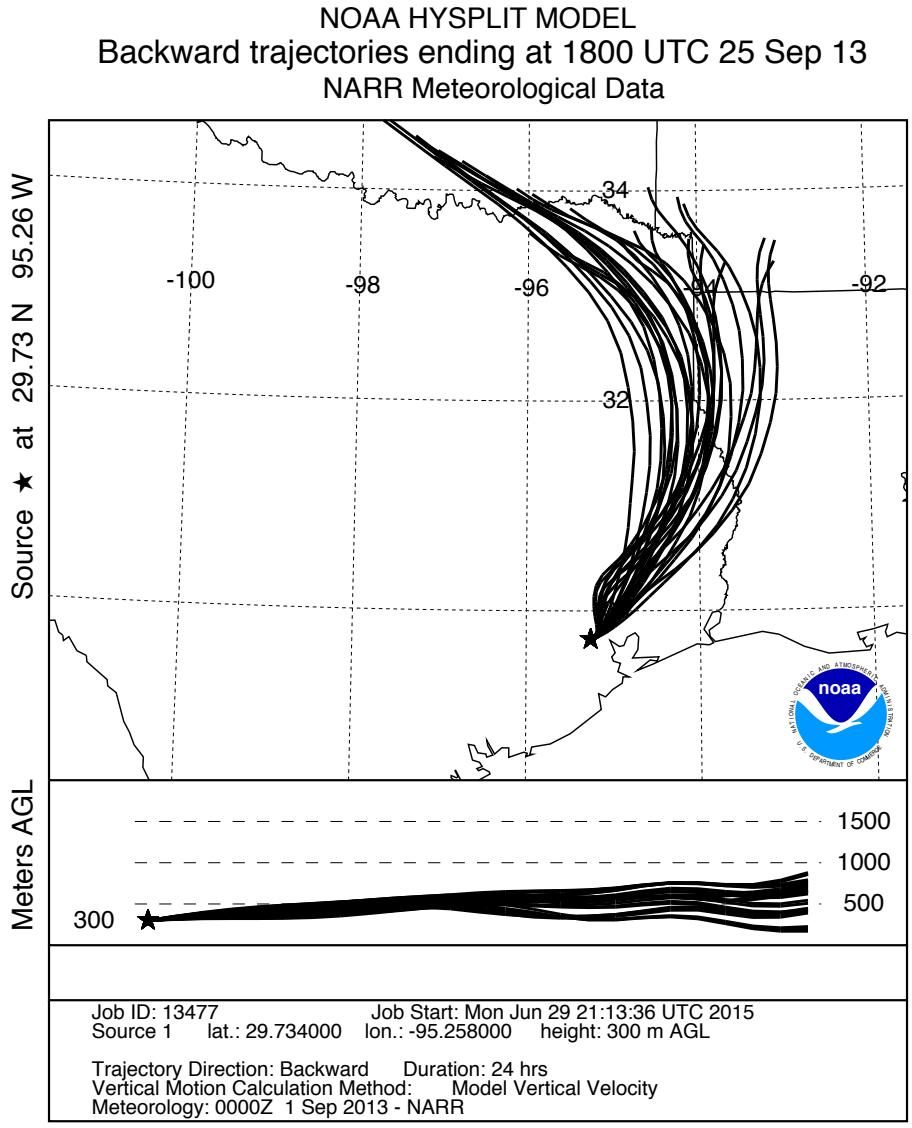


Figure A.1. Ensemble back-trajectory run for the Houston/Galveston/Brazoria area on August 25, 2013.

A.1.3 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.

A.1.4 Baseline GAMs (gam01_baseline)

A.1.4.1 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 A.4 below. The automated process to remove insignificant predictors was not used for these fits.

Table A.4. 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, *morning_ws*)
Afternoon (1–4 pm) average wind speed (m/s, *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.*)

A.1.4.2 Results

To illustrate the results, we discuss the baseline GAM fits for HGB in detail. Similar plots for all urban areas are contained in the deliverable as described in Section A.2.6. Figure A.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 A.4. 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 HGB. 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 A.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 A.1.6 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 A.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 HGB 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 A.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.

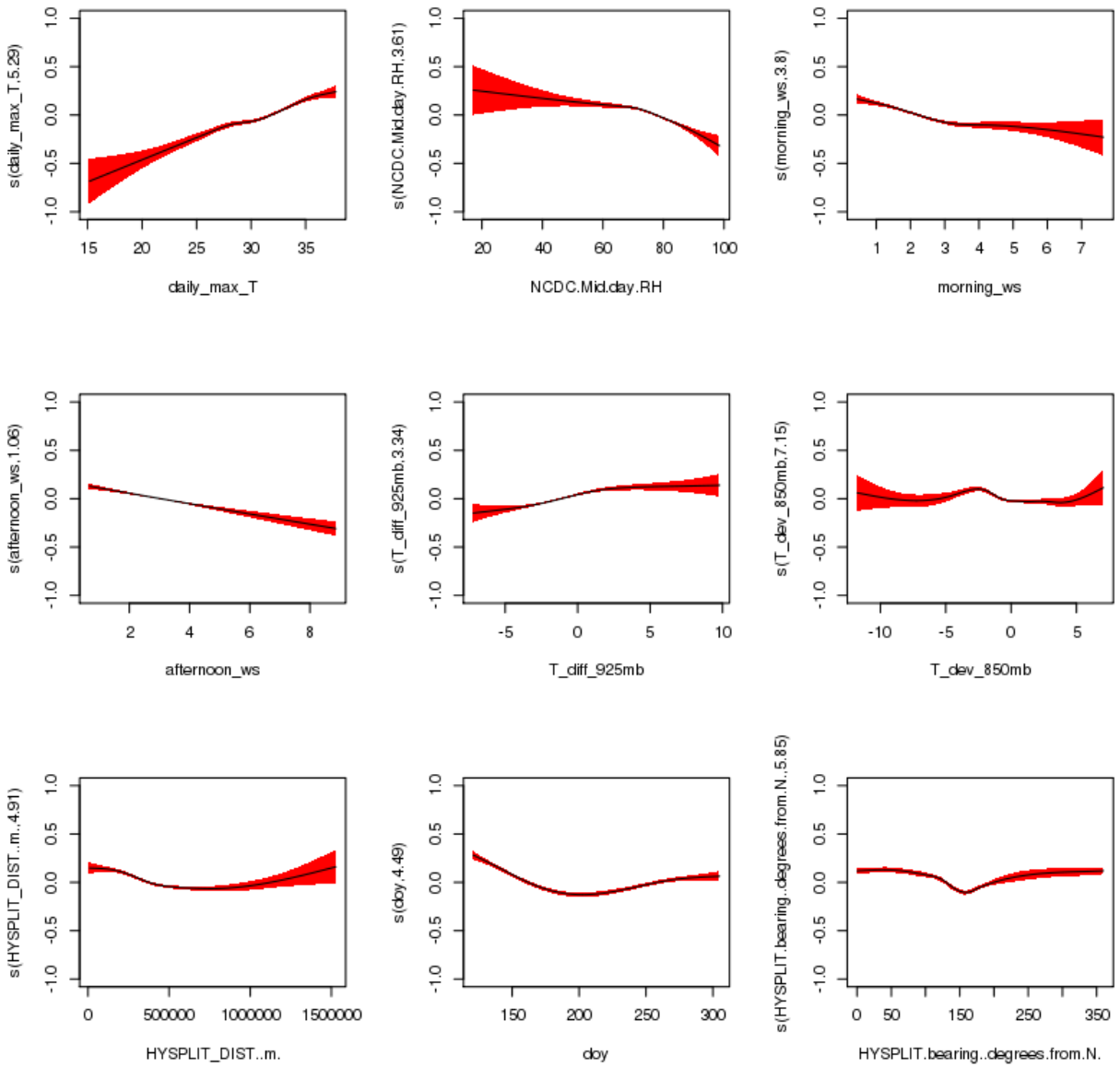


Figure A.2. Smooth functions for the baseline GAM (gam01_baseline) fit to HGB MDA8 O_3 data. The y-axis scale is the scale of the "linear predictor", i.e. the deviation of the natural logarithm of the MDA8 O_3 in ppbv from its mean value.

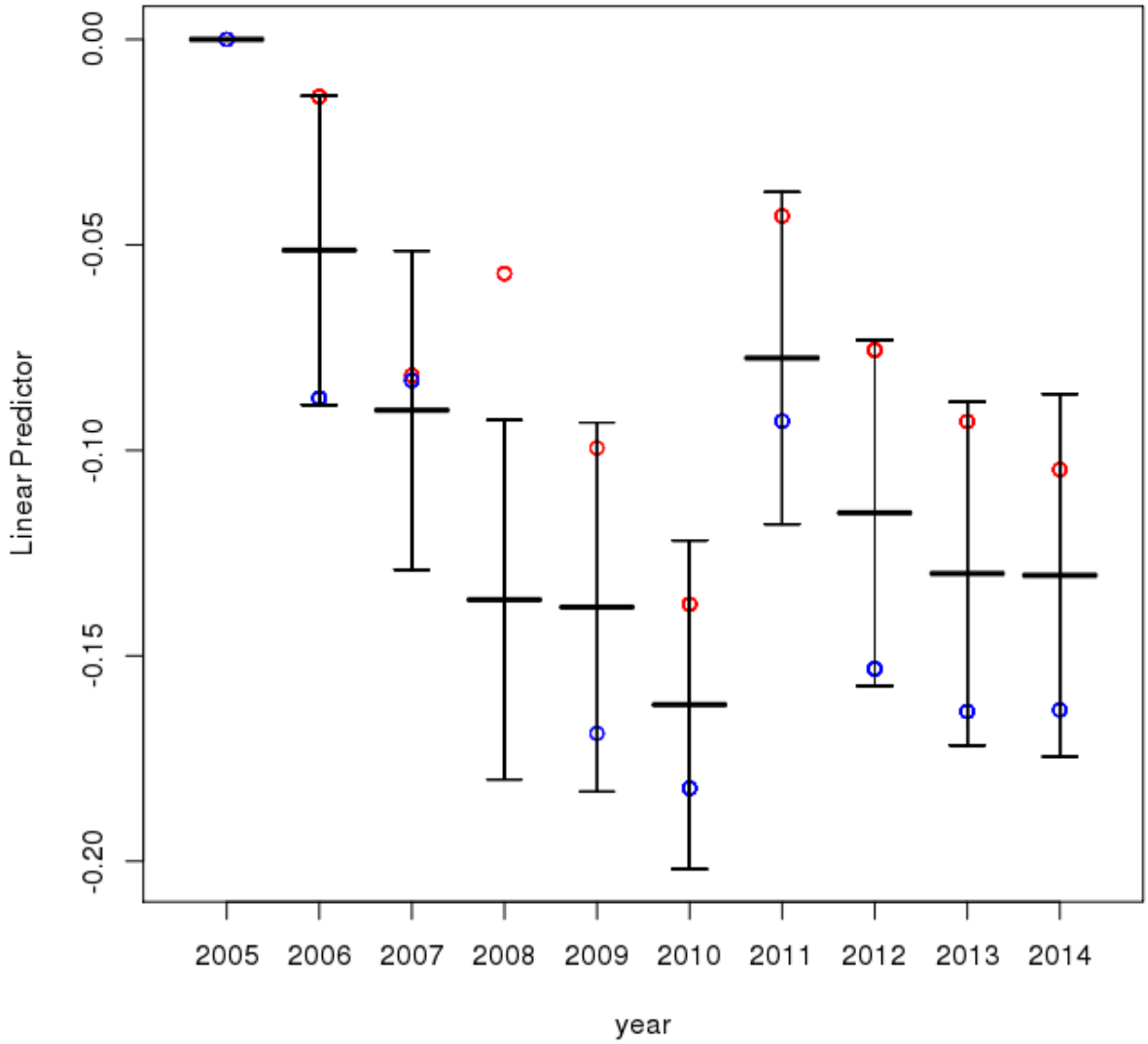


Figure A.3. Year-to-year deviations from 2005 for the baseline GAM (gam01_baseline) fit to HGB MDA8 O_3 data. The y-axis scale is the scale of the "linear predictor", i.e. the deviation of the natural logarithm of the MDA8 O_3 in ppbv from its mean value. The black center bar is the mean value while 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 A.1.6.

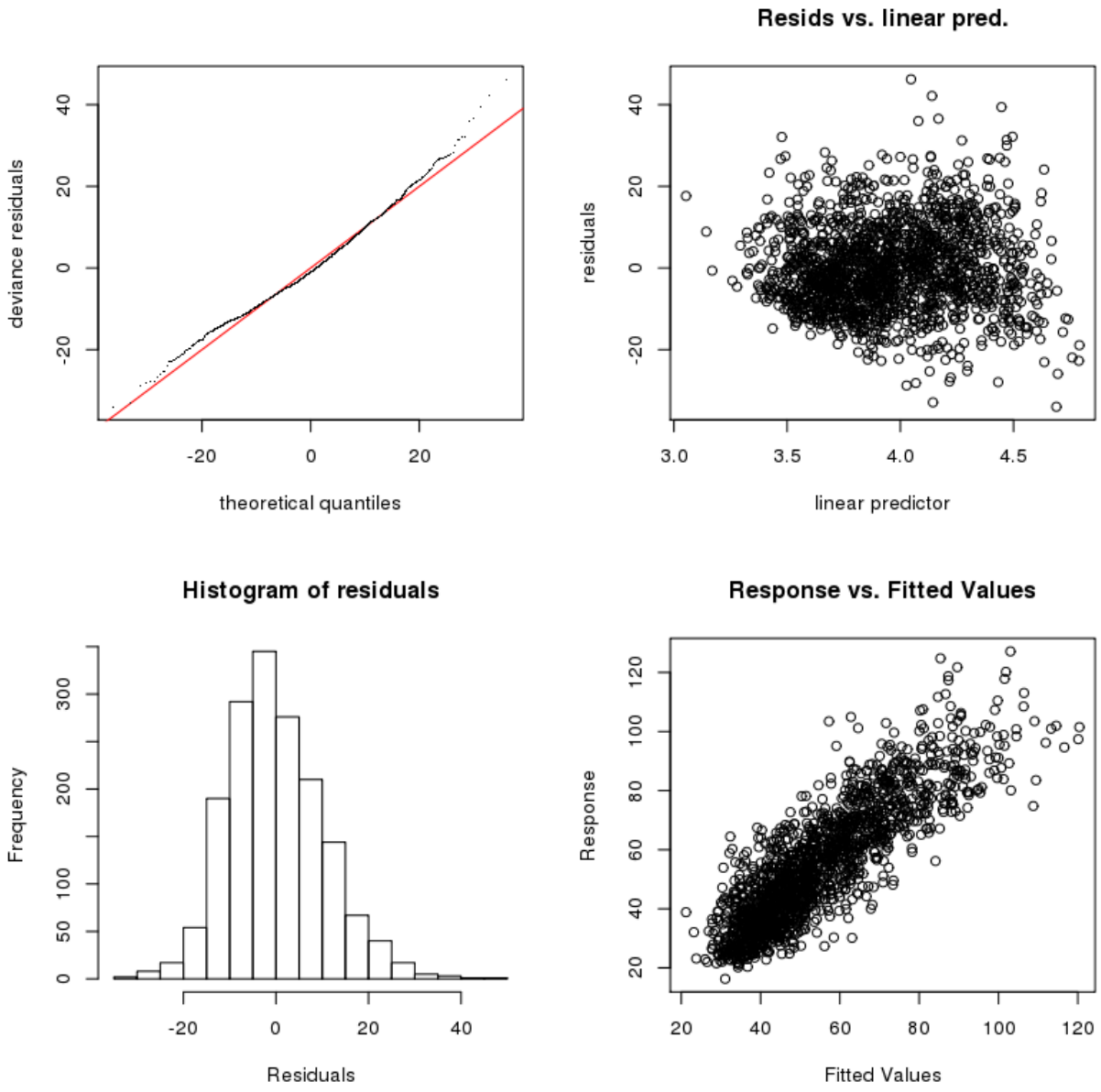


Figure A.4. GAM evaluation plots for the baseline GAM (gam01_baseline) fit to HGB MDA8 O_3 data.

Figure A.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 A.4 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 HGB 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 A.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 A.1.6 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 A.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 A.5 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% (SA) and 73.9% (HGB), 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% (BPA) and 37.8% (HGB).

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 A.5. 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 A.1.6.

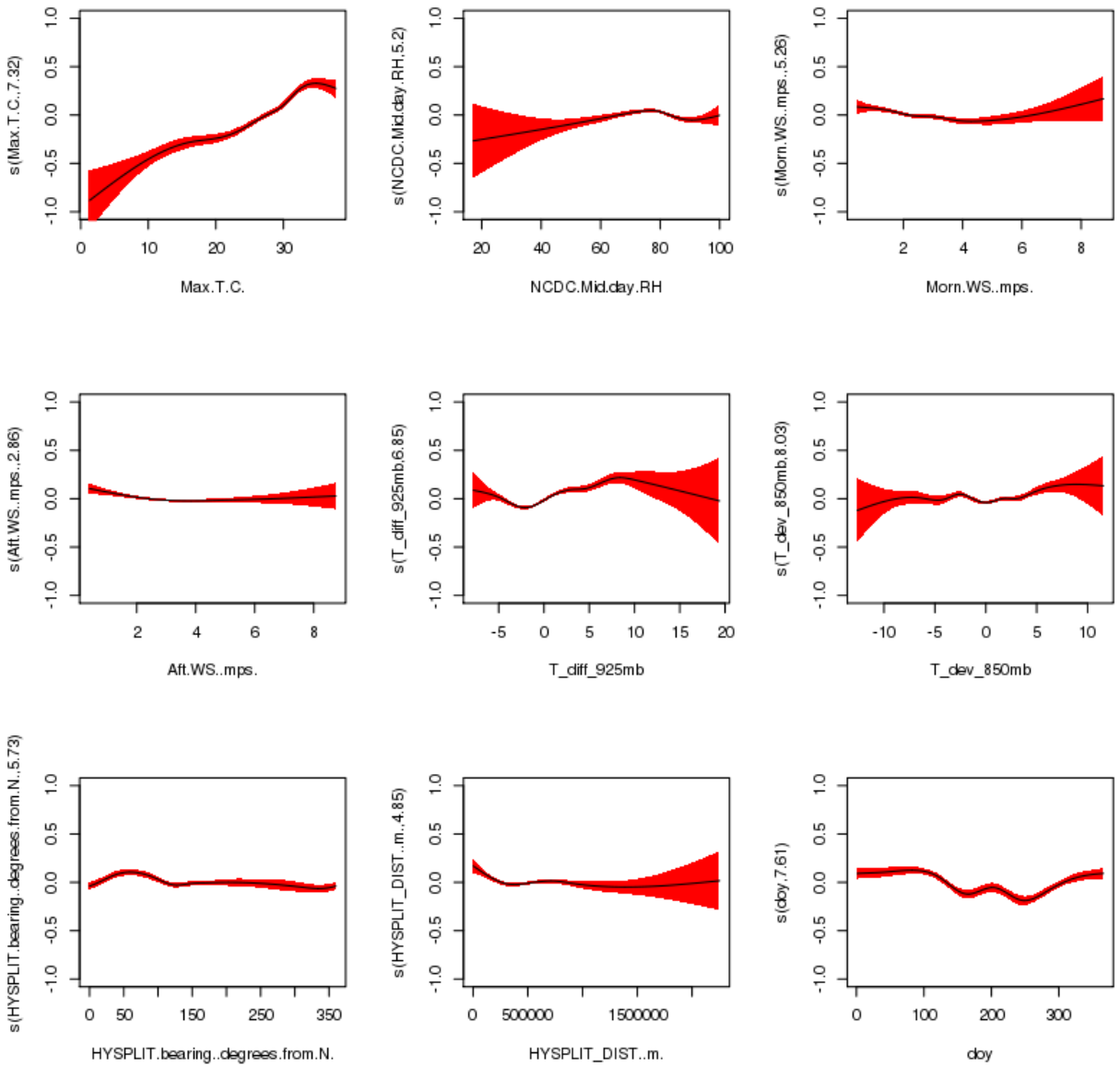


Figure A.5. Smooth functions fit for the baseline GAM (gam01_baseline) fit to HGB daily average $PM_{2.5}$ data. The y-axis scale is the scale of the "linear predictor", i.e. the deviation of the natural logarithm of the daily average $PM_{2.5}$ in µg m^{-3} from its mean value.

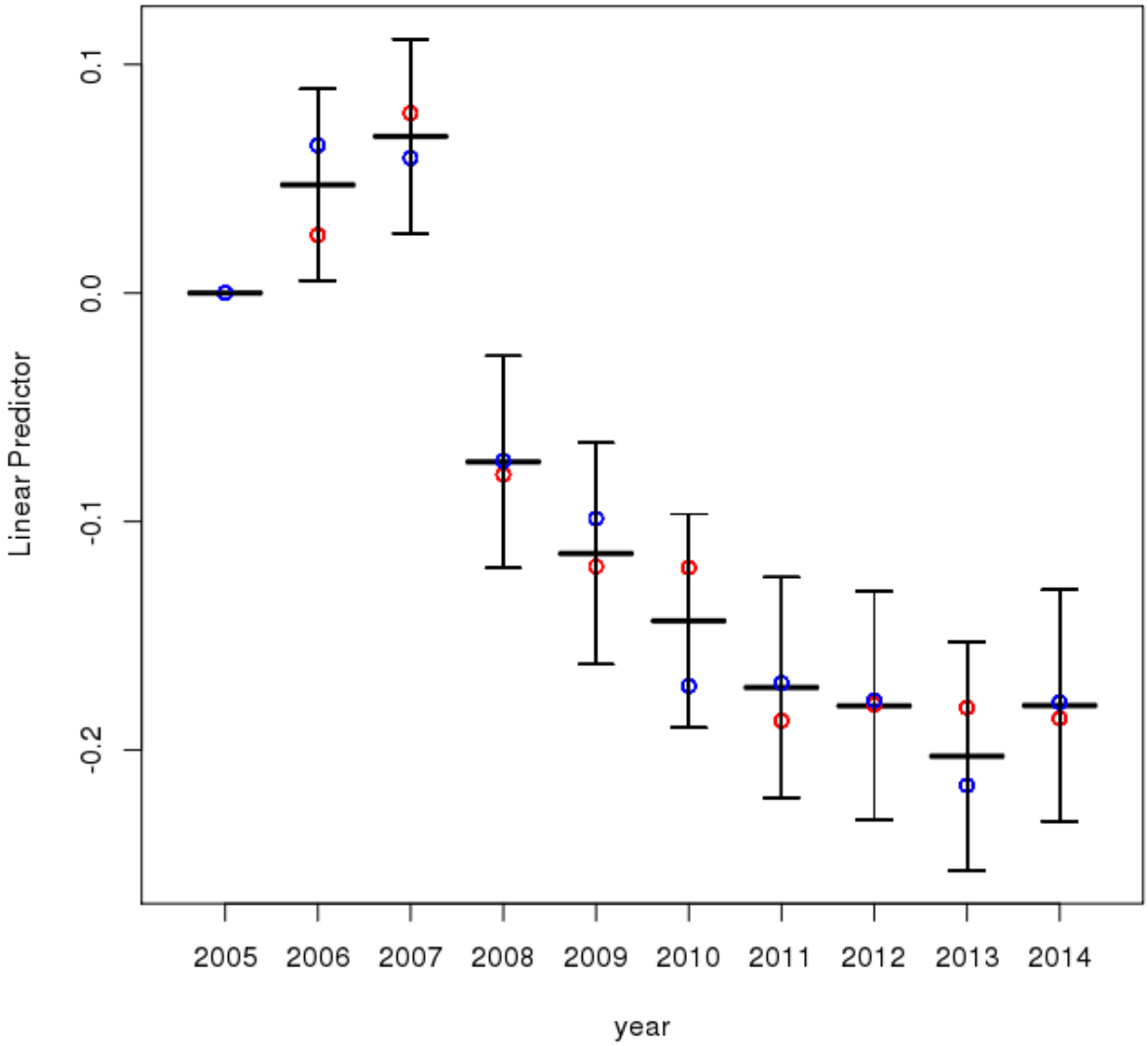


Figure A.6. Year-to-year deviations from 2005 for the baseline GAM (gam01_baseline) fit to HGB daily average $PM_{2.5}$ data. The y-axis scale is the scale of the "linear predictor", i.e. the deviation of the daily average $PM_{2.5}$ in μg m^{-3} from its mean value. The black center bar is the mean 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 A.1.6.

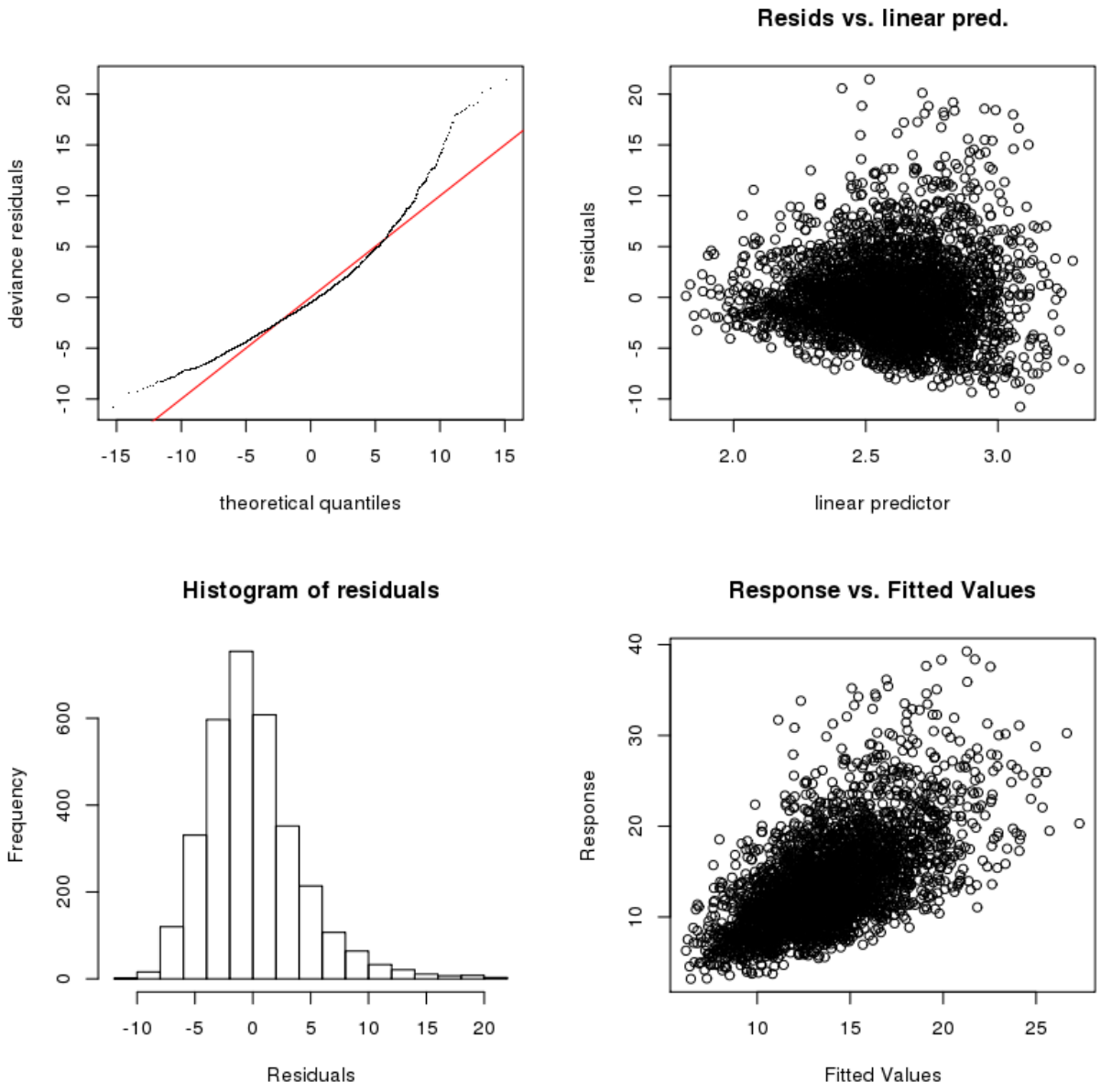


Figure A.7. GAM evaluation plots for baseline GAM (gam01_baseline) fit to HGB MDA8 O_3 data.

Table A.6 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 greater than 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 approximately 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 approximately 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 approximately 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 greater than approximately 30 ^oC.
- The fits for mid-day RH are very uncertain at low (less than 40%) and high (greater than 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 approximately 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 approximately 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 approximately 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 approximately 200 Julian days (July).

Table A.5. 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 A.6. 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*

A.1.5 Extended GAMs (gam02_extended and gam03_extended)

A.1.5.1 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 A.1.4 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 A.1.1, as well as the HYSPLIT back-trajectory endpoints described in Section A.1.2. 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 A.2.4.2, and predictors in those files are listed 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 greater than 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 A.7 while the predictors for maximum daily average $PM_{2.5}$ are listed in Table A.8. These predictors were used to fit the "large" extended GAMs (gam02_extended). These fits did use the automated selection procedure described in Section A.1.3 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 Table A.7 and Table A.8.. 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 A.7. 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)	*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 A.8. 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)	*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**

A.1.5.2 Results

Table A.9 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 Figure 23 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 A.1.4, 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 (DFW and HGB) 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 A.9. 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 A.10. 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 A.1.4, we discuss the small extended GAMs for HGB in detail to illustrate the results. Similar plots for all urban areas are contained in the deliverable as described in Section A.2.6. Figure A.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 HGB. 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 A.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 A.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 A.10. These plots indicate a good fit, as the model residuals are roughly normally distributed and show no trend versus predicted value.

Figure A.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 HGB 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 A.12, and are very similar to the results for the baseline case shown in Figure A.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 A.1.6 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 A.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 A.11 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 SA or ARR.

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 greater than 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 approximately 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 approximately 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 approximately 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 greater than approximately 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 (approximately 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 approximately 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 approximately 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 approximately 150° (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 approximately 200 Julian days (July).

We also compared the functional forms in the extended GAMs to those of the baseline GAMS described in Section A.1.4. 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 A.11. 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*

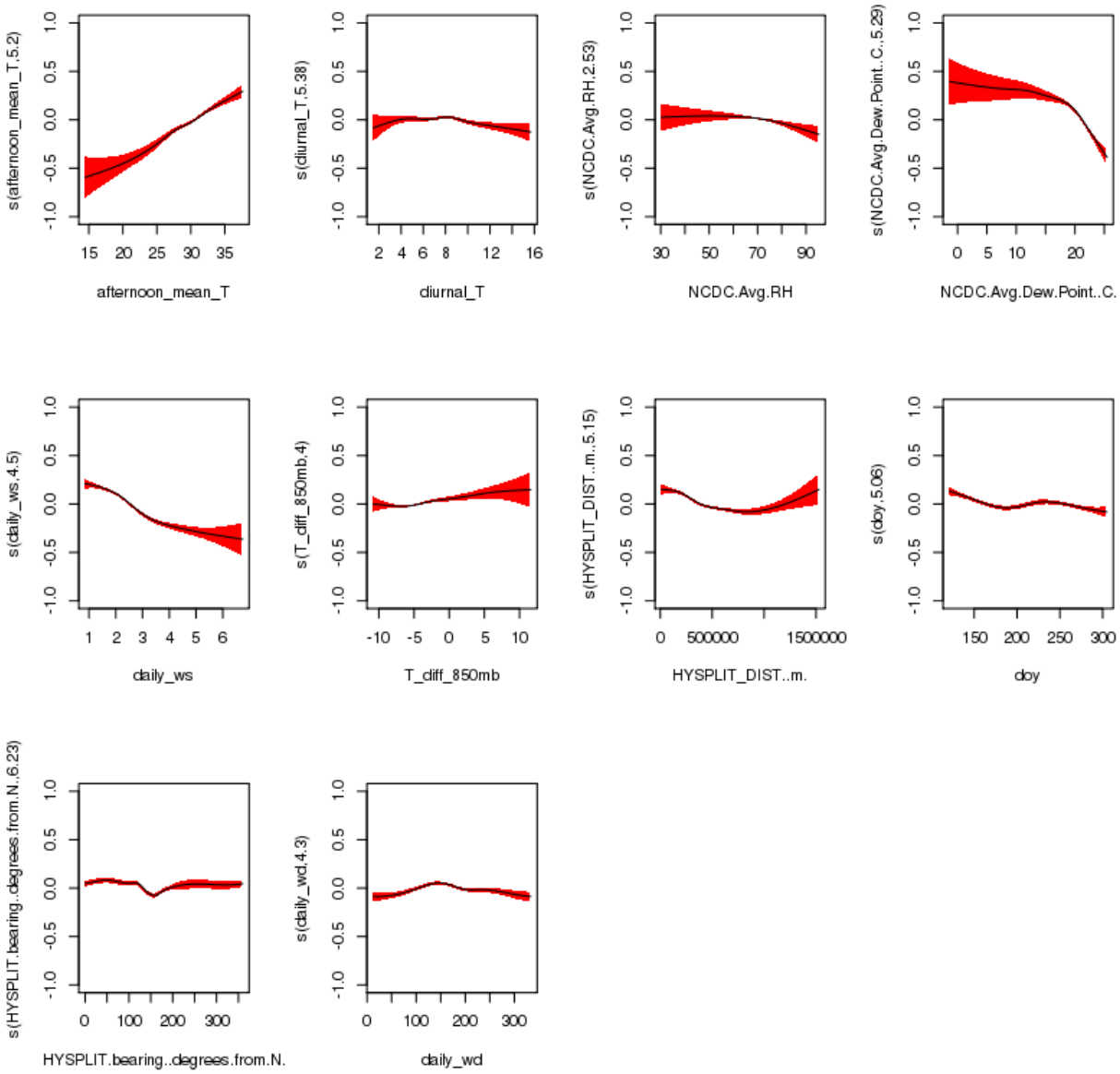


Figure A.8. Smooth functions for the small extended GAM (gam03_extended) fit to HGB MDA8 O_3 data.

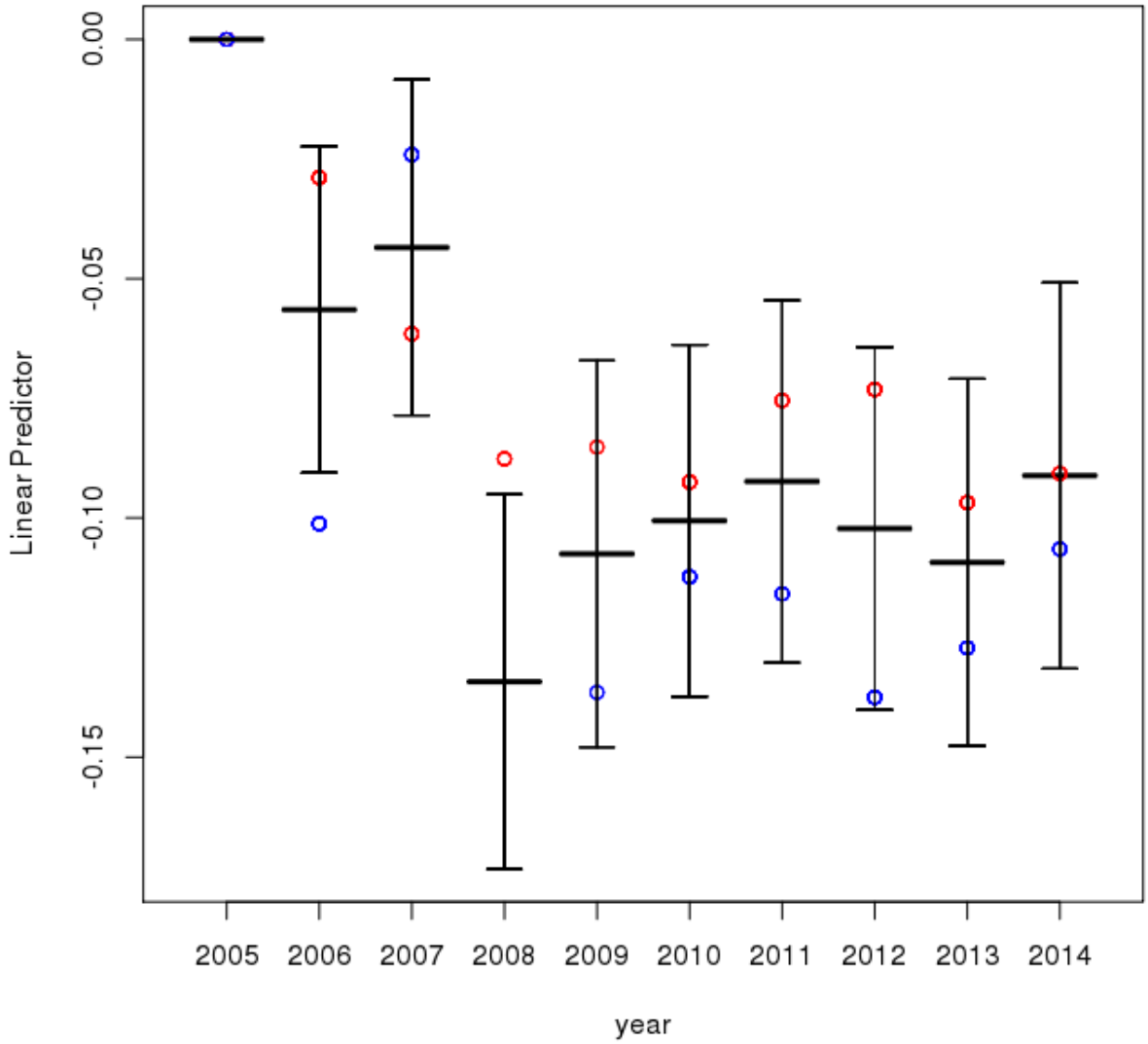


Figure A.9. Year-to-year deviations from 2005 for the small extended GAM (gam03_extended) fit to HGB MDA8 O_3 data.

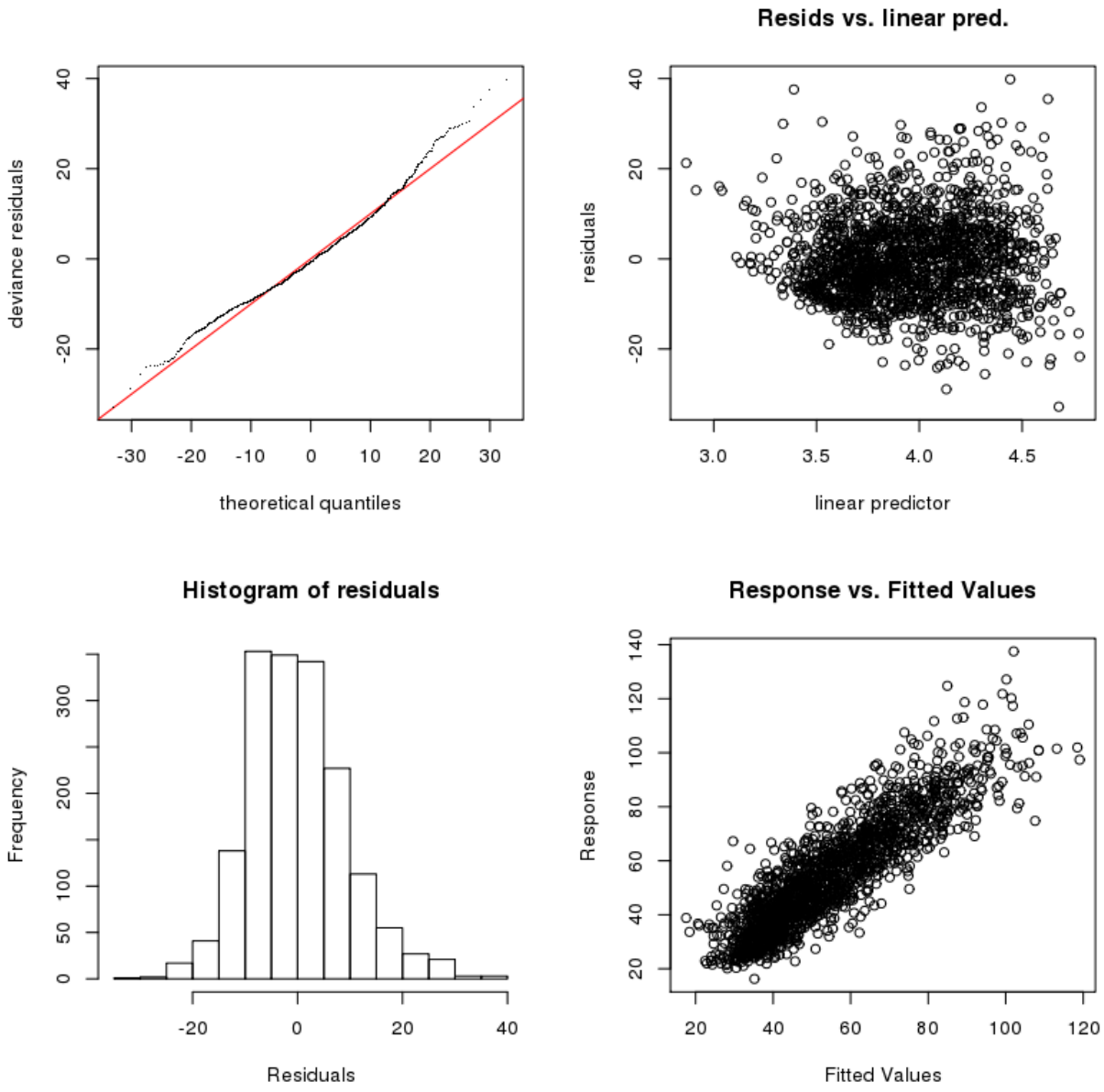


Figure A.10. GAM evaluation plots for the small extended GAM (gam03_extended) fit to HGB MDA8 O_3 data.

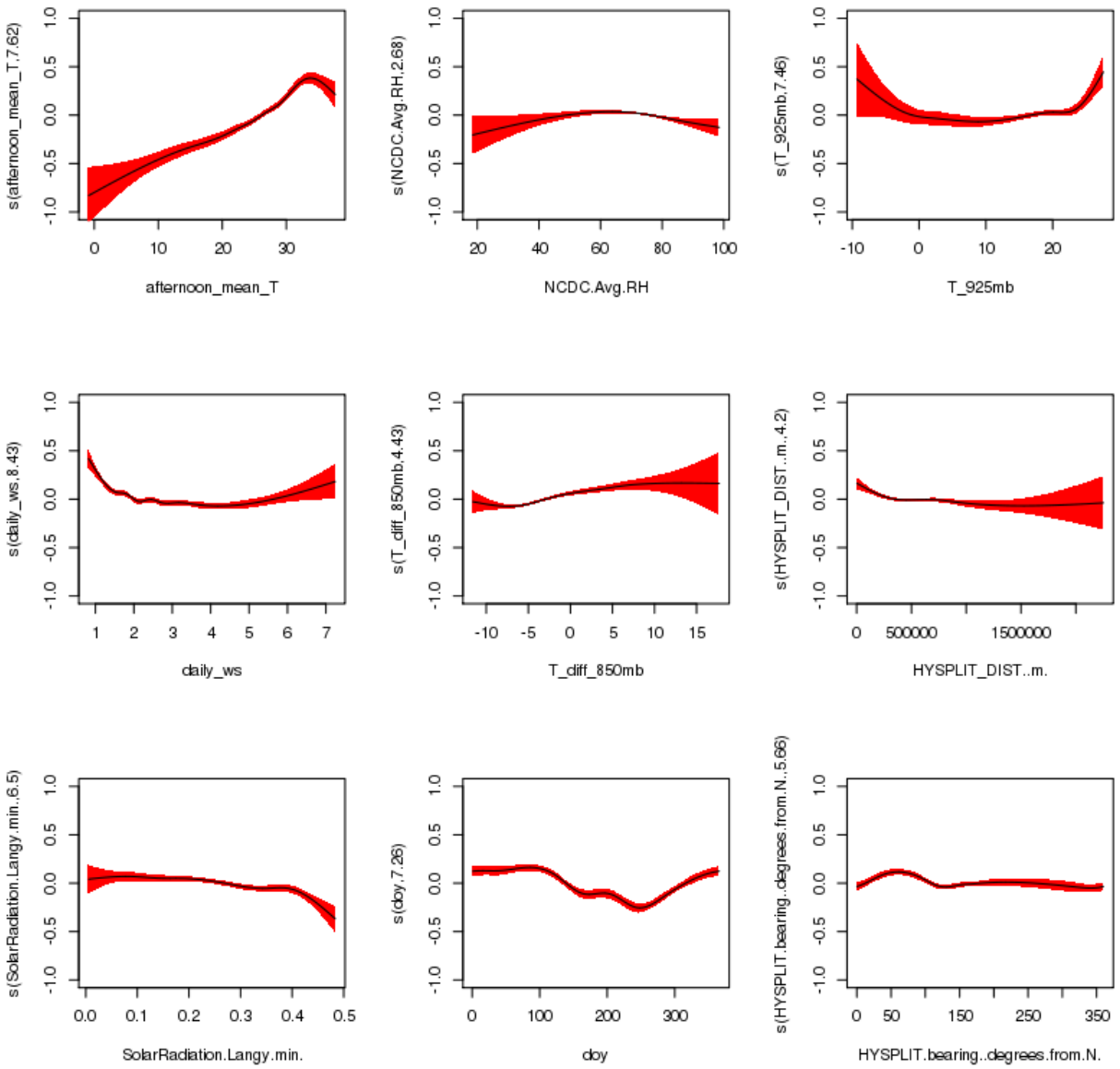


Figure A.11. Smooth functions for the small extended GAM (gam03_extended) fit to HGB daily average $PM_{2.5}$ data.

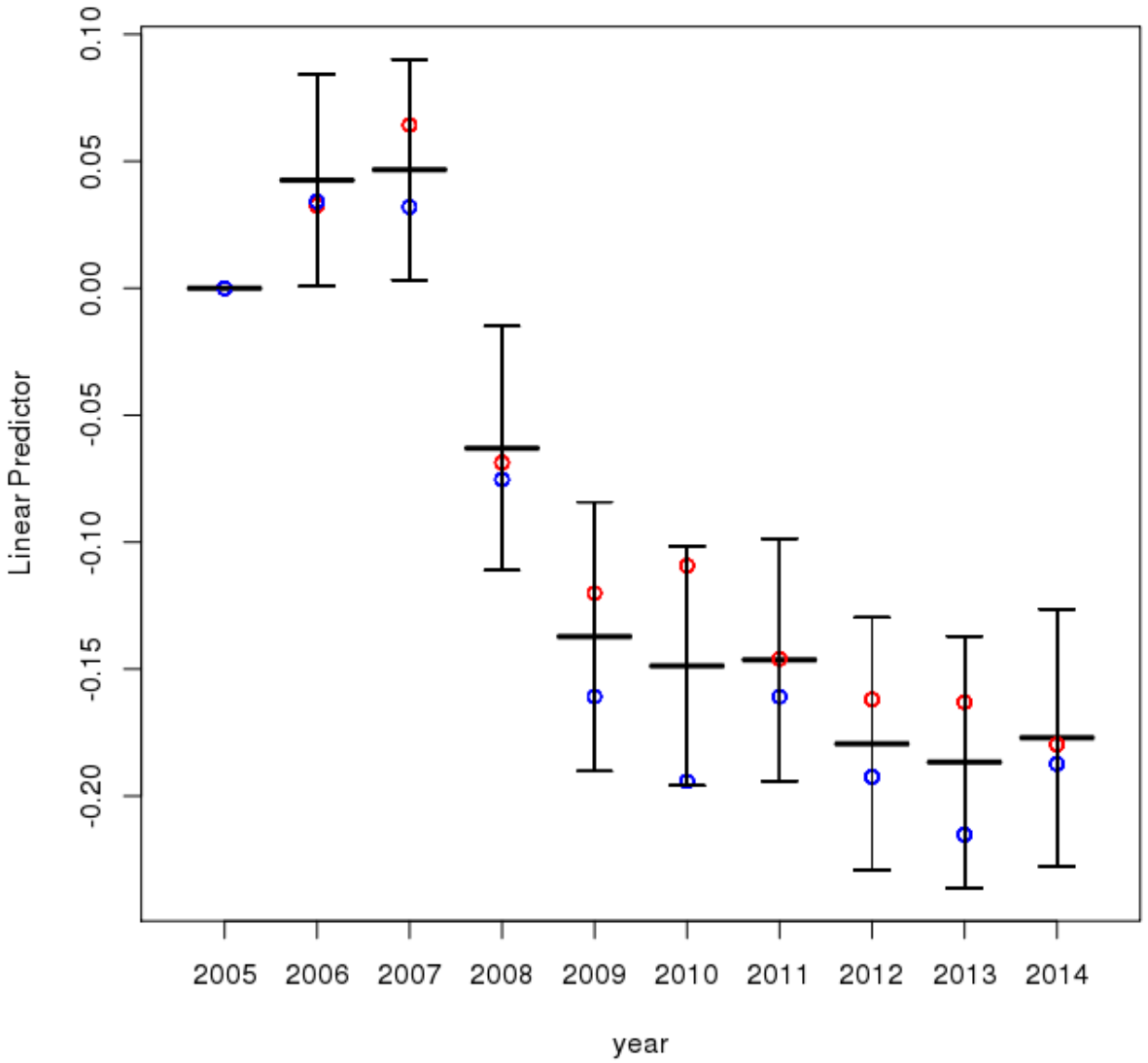


Figure A.12. Year-to-year deviations from 2005 for the small extended GAM (gam03_extended) fit to HGB daily average $PM_{2.5}$ data.

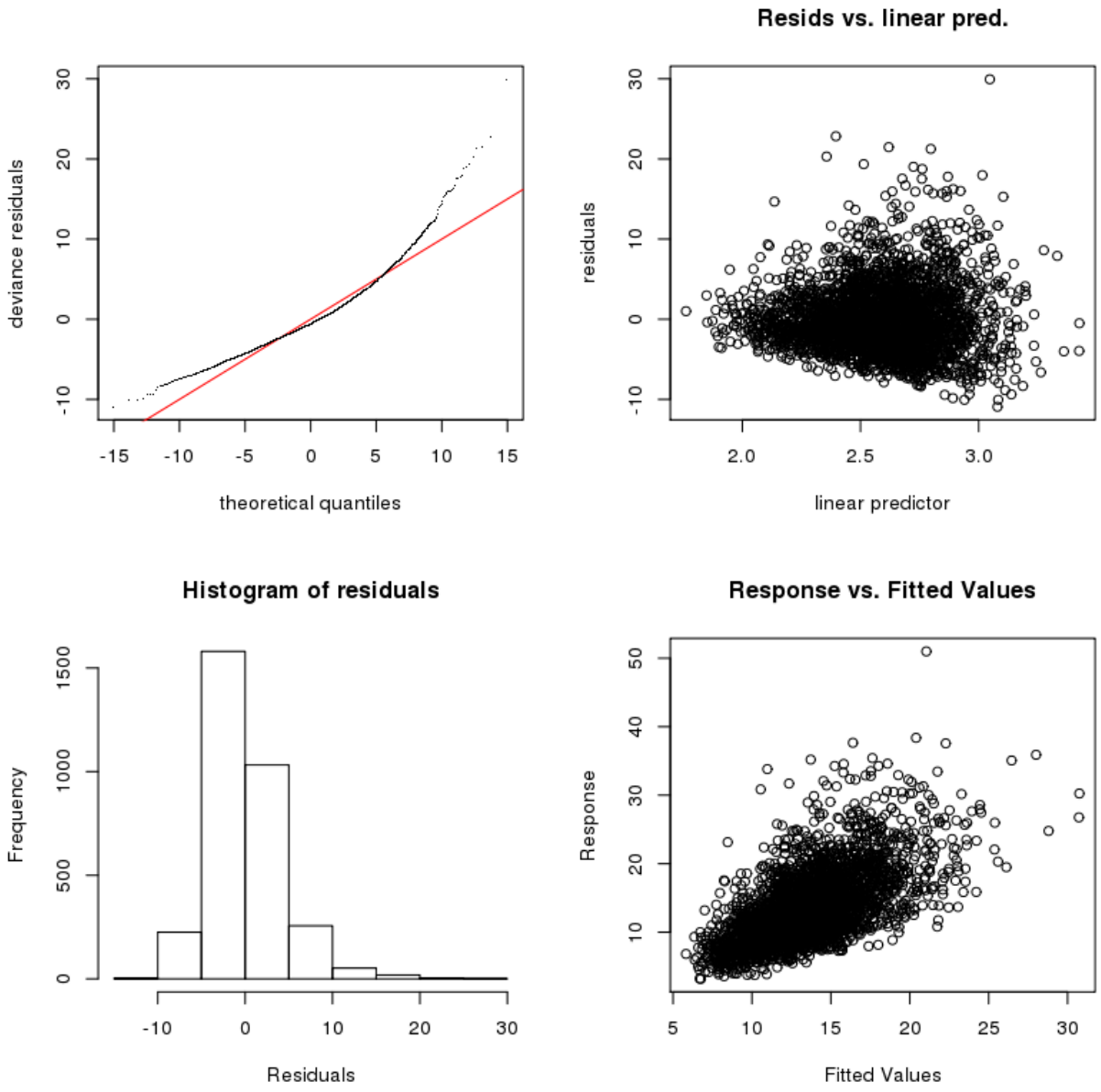


Figure A.13. GAM evaluation plots for the small extended GAM (gam03_extended) fit to HGB daily average $PM_{2.5}$ data.

A.1.6 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 A.14 shows scatterplots of the GAM-predicted (x-axis) versus the measured (y-axis) values of maximum daily average $PM_{2.5}$ for the HGB 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 A.12, 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 (less than 1 ppbv for O_3 and less than 0.25 $\mu g\ m^{-3}$ 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 A.15 shows the HGB 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 A.15 for all GAMs and their terms are contained in the deliverable, as described in Section A.2.6. Other "suspicious" functional forms for $PM_{2.5}$ and O_3 in the gam03_extended fits are listed in Table A.13, 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.

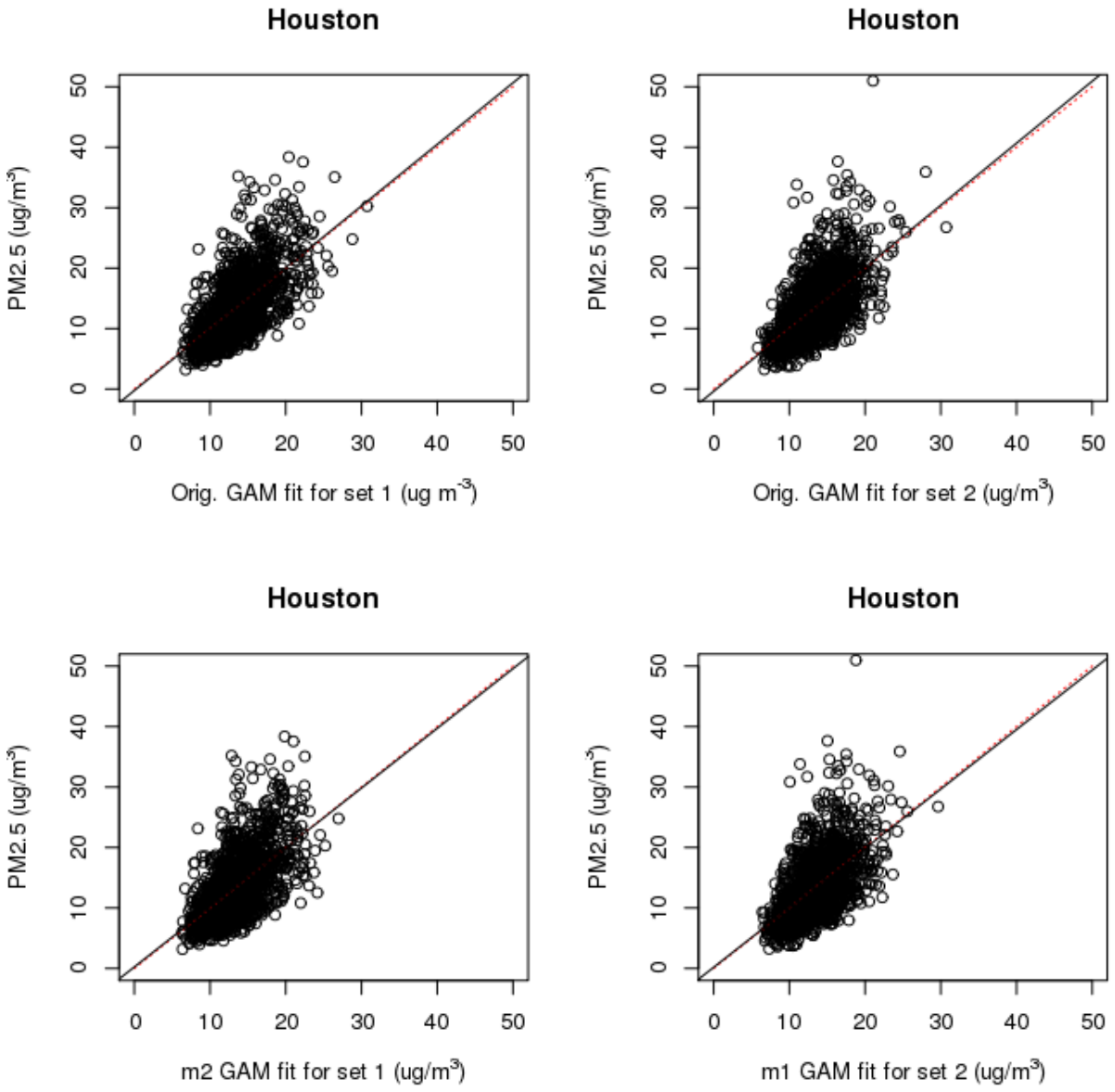


Figure A.14. Scatterplots for the GAM-predicted (x-axis) versus the measured (y-axis) values of maximum daily average $PM_{2.5}$ for the Houston/Galveston/Brazoria area using gam03_extended. The top row uses m_{tot} to predict the first (left) and second (right) of the randomly distributed halves of 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.

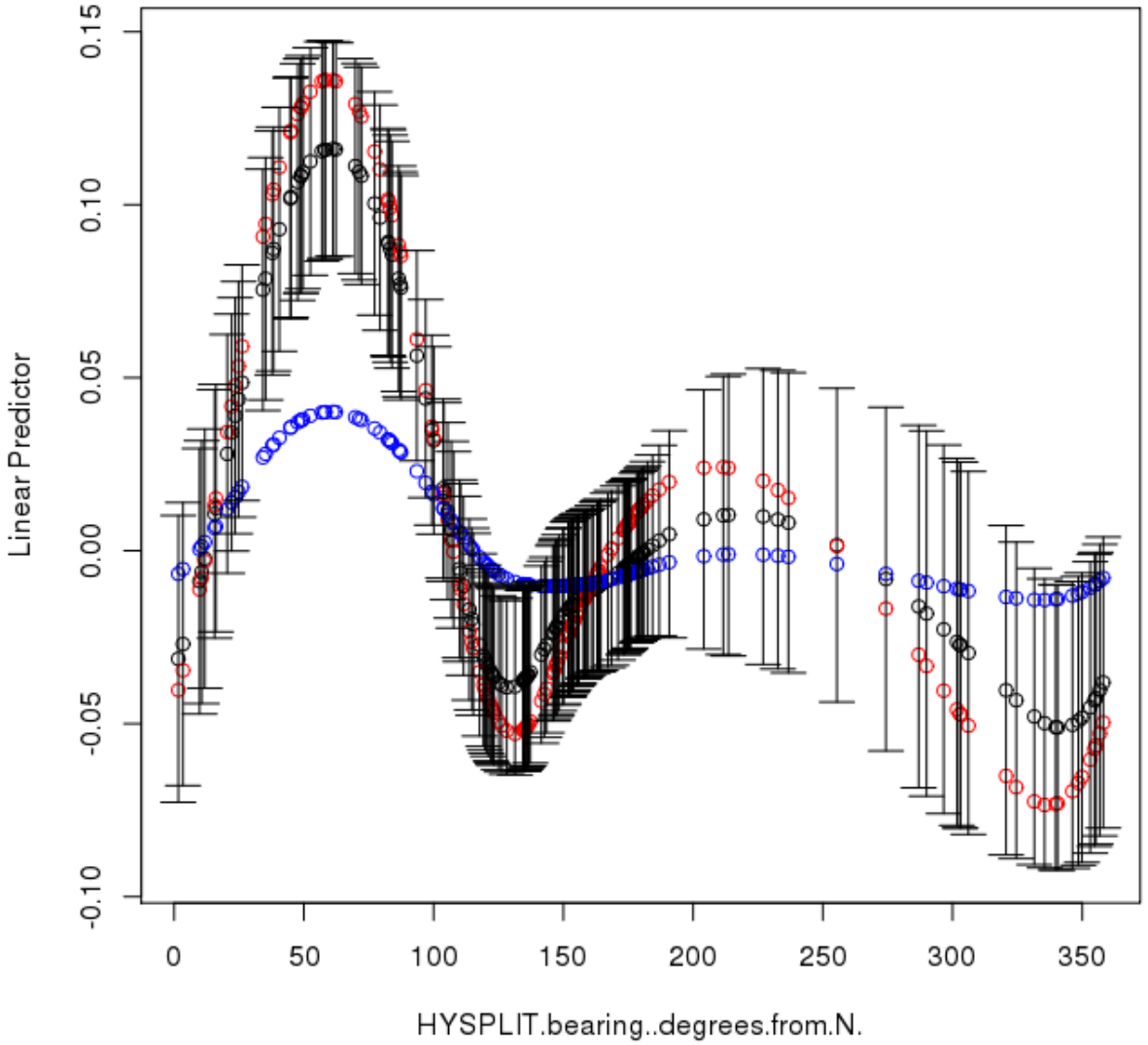


Figure A.15. Houston/Galveston/Brazoria fits for maximum daily average $PM_{2.5}$ versus HYSPLIT back trajectory bearing for m_{tot} (black with error bars), m_1 (red) and m_2 (blue) for gam03_extended. Predicted values for 200 randomly selected datapoints are plotted.

Table A.12. Cross-validation root-mean-square (RMS) results for gam03_extended.

Urban Area	MDA8 O_3 (ppbv)				Daily Average $PM_{2.5}$ ($\mu g\ m^{-3}$)			
	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 A.13. "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*

A.2 File Descriptions

This section describes all of the files included in the deliverable. Figure A.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 A.2.3.1 to A.2.4. Figure A.17 shows the scripts that use the CSV file produced at the end of Figure A.16 to produce and evaluate the GAMs. These scripts and the output files produced are described in Sections A.2.5 and A.2.6, respectively.

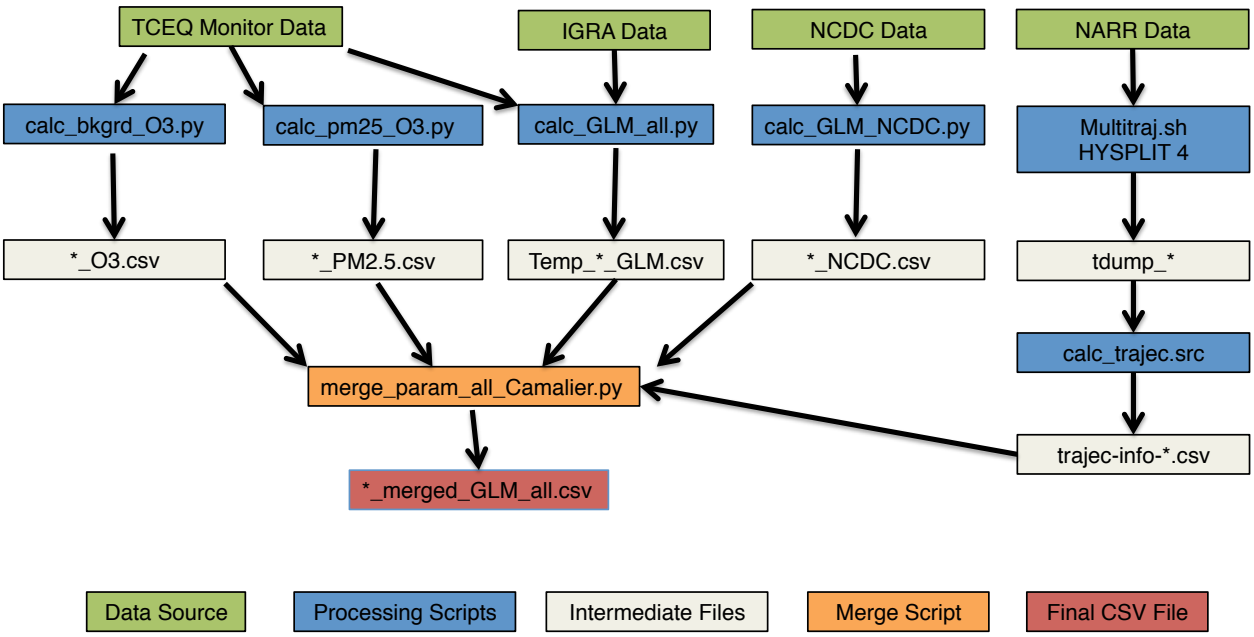


Figure A.16. Flow chart showing the processing from the original data sources (green boxes) to the final CSV file (red box) that is used as input for the GAM fitting scripts.

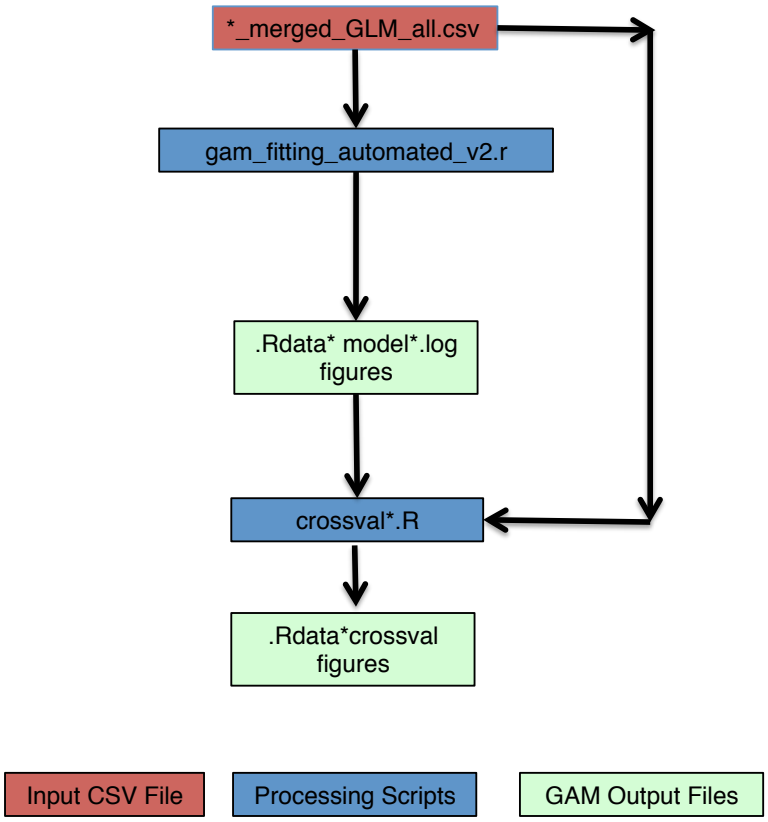


Figure A.17. Flow chart showing the processing from the input CSV file generated at the end of Figure A.16 (red box) to the GAM output files (light green box).

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.

A.2.1 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.

A.2.1.1 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 A.2, 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.

A.2.1.2 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 A.3. The raw data file contains daily data from 2005-2014. Missing data is indicated by ***.

A.2.2 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 A.1.1 and A.2.1. The selected background sites are listed in Table A.14 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 GAM 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 A.1.1. These background sites are listed in Table A.14 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 GAM parameters. The outputs of this script were previously supplied to TCEQ as Deliverable 3.1 (Appendix B. Technical Memo: Estimating Background O_3 and $PM_{2.5}$).

- *./scripts/calc_GLM_all.py* : This script reads in TCEQ monitor site and IGRA (upper atmosphere) measurements to derive daily GAM parameters described in the script itself. It performs all the necessary conversions (ex. Fahrenheit to Celsius, mph to m/s) 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 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. 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 A.2.5.

Table A.14. 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

A.2.3 HYSPLIT

A.2.3.1 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 A.1.2. 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.

A.2.3.2 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

A.2.3.3 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. 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.
- *./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 A.2.2).

A.2.4 Processed Input Data Files in CSV Format (*./csv_files/*)

A.2.4.1 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 A.2.2 (*./scripts/calc_GLM_NCDC.py* and *./scripts/calc_GLM_all.py* respectively). They contain daily GAM values for all urban locations from 2005-2014 and are used as input by *./scripts/merge_param_all_Camalier.py*.

A.2.4.2 Final CSV Files (./csv_files/final_files/)

These files are created by *./scripts/merge_param_all_Camalier.py* (Section A.2.2), which combines all daily meteorological predictors with the O_3 and $PM_{2.5}$ concentrations for each location. The file 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 A.2.5.

A.2.5 GAM scripts (./full_gam_fits/)

A.2.5.1 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 A.1.5.1. 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 $PM_{2.5}$.

A.2.5.2 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 A.2.6):
    - 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.

A.2.5.3 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.

A.2.6 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 A.1.3.
- *plot_o3gam03_extended_Houston_smoothfunc-noresid.png* : A figure showing the smooth functional fits for the GAM, as in Figure A.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 A.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 A.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 A.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 A.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 A.1.6.

A.3 Quality Assurance Steps

In addition to the analyses described in Section A.1.3, other quality assurance checks were made. All scripts used in this project were inspected by team 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 A.3.1 and A.3.2, respectively.

A.3.1 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 A.1.4.2 and A.1.5.2, 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 A.1.4.2, 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 A.1.2, 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 approximately 100 km and differences in bearing of less than approximately 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 A.1.6, 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 A.1.6, 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.

A.3.2 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 A.1.4 and A.1.5. The software versions and computers used to run the scripts supplied in the deliverable are documented in the beginning of Section A.2.

- *The equations on which the model is based*

  The main GAM equation is given in Section A.1.3. 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 A.1. 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 A.16 *and* Figure A.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 A.2.2, A.2.3, and A.2.5.

- *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 A.2.4. The sources of this data are described in Section A.1.1.

- *A copy of the source code*

Copies of all scripts developed in this project are contained in the deliverable, as described in Sections A.2.2 and A.2.5.

- *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 Section A.2.4. 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 A.2.3.2. The post-processed distance and bearing outputs are contained in the intermediate CSV files described in Section A.2.3.3 and in the final CSV files described in Section A.2.4.2.
- *Output of model runs and interpretation*
  Section A.2.6 describes the output files from our GAM fits and cross-validation analysis contained in the deliverable. These results are discussed and interpreted in Sections A.1.4, A.1.5, and A.1.6.
- *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 A.2.4.2. The units of the variables much match those given in Table A.4 (gam01_baseline), Table A.7 (gam02_extended), and Table A.8 (gam03_extended). The data processing scripts described in Section A.2.2 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 A.2. Section A.2.6 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 A.1.3. The verification of the GAMs was performed via the two-fold cross-validation described in Section A.1.6.

  The evaluation of the HYSPLIT back-trajectories is described in Section A.1.2. The GAMs were evaluated as described in Sections A.1.6, as well as by addressing the quality assurance questions in Section A.3.1.

Appendix B. Technical Memo: Estimating Background O_3 and $PM_{2.5}$

B.1 Introduction

This appendix documents the files provided to TCEQ to complete Deliverable 3.1 of Work Order No. 582-15-54118-01. Section B.2 briefly outlines the technical approach used to prepare the files in the deliverable (provided via email to Erik Gribbin of TCEQ as a gzipped tar file: p1952_deliverable_3_1_R1_1.tar.gz) and Section B.3 describes the format of the files. Section B.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.

B.2 Technical Approach

As described in the Work Plan, our approach follows the TCEQ method described in Berlin et al. (2013). This method requires: the selection of background sites; the calculation of the maximum daily 8-hour average (MDA8) for ozone (O_3) and the daily average of fine particulate matter ($PM_{2.5}$) at each site; estimating a preliminary background value as the lowest of the valid values for the background sites; and then further investigations to ensure the values are appropriate background estimates. These steps are described in detail below.

B.2.1 Selection of Background Sites

The initial data for our analysis was provided by Erik Gribbin of TCEQ, which consisted of hourly-average measurements of O_3 and $PM_{2.5}$ at several sampling sites surrounding the urban areas of interest. After consultation with TCEQ, we selected "background" monitor sites near the edge of each urban area. These background sites were chosen to be at a significant distance from major pollutant emission sources. The AQS site numbers for the selected background sites for each urban area are given in Table B.1. Note that for the ARR and TLM areas, most or all of the available urban sites are considered potential "background" sites due to the limited number of sampling sites available.

Table B.1. 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

To calculate the background MDA8 O_3 for the State of Texas as a whole, we used two approaches, the first using data from TCEQ sites near the Texas border, and the second using data from sites in the US EPA Clean Air Status and Trends Network (CASTNET)[4]. The CASTNet sites used to calculate Texas background O_3 are listed in Table 3; a csv file (CASTNet_site_info.csv) with the latitudes, longitudes, and elevations of the CASTNet sites is included in the deliverable.

To calculate the background daily average $PM_{2.5}$ for the State of Texas, we used data from sites in the Interagency Monitoring of Protected Visual Environments (IMPROVE)[5] network near the Texas border, as TCEQ sites near the Texas border rarely made $PM_{2.5}$ measurements. The IMPROVE sites used to calculate Texas background O_3 are listed in Table B.2; a csv file (IMPROVELocTable.csv) with the latitudes, longitudes, and elevations of the IMPROVE sites is included in the deliverable.

Table B.2. Sites used to calculate background O_3 and $PM_{2.5}$ for the State of Texas as a whole.

Pollutant (Network)	# of Background Sites	IDs of Background Sites
O_3 (TCEQ)	23	484790017, 484790016, 484790313, 482150043, 480610006, 482730314, 483550025, 482450101, 481675005, 480391003, 482030002, 480370004, 482311006, 483670081, 480650007, 480650004, 480650005, 481351014, 481350003, 481410058, 800060003, 481410029, 481410057
O_3 (CASTNet)	10	CHA467, PET427, MEV405, CHE185, CAD150, CVL151, SUM156, EVE419, PAL190, BBE401
$PM_{2.5}$ (IMPROVE)	12	BOAP1, SAAN1, WHIT1, GUMO1, SACR1, BIBE1, ELLI1, WIMO1, CACR1, SIKE1, HOUS1, BRET1

[4] U.S. Environmental Protection Agency Clean Air Markets Division, *Clean Air Status and and Trends Network (CASTNET),* Table OZONE_8HR_DMAX, last updated 2015-04-06. Available at www.epa.gov/castnet. Accessed 2015-04-09.

[5] *Interagency Monitoring of Protected Visual Environments (IMPROVE) network*, Table EPA PM2.5 Mass FRM – Daily, Available at http://vista.cira.colostate.edu/improve/Data/IMPROVE/improve_data.htm. Accessed 2015-04-09.

B.2.2 Calculation of MDA8 O_3 and daily average $PM_{2.5}$ Values for Each Site

We developed a python script (calc_bkgrd_ozone.py)[6] 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:

5. 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.
6. The largest of each of the calculated 8-hour averages in a day was selected as the MDA8 for that day.

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:

5. If more than one $PM_{2.5}$ instrument was active for a site, the reported hourly values were averaged.
6. 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.

For the background values for the State of Texas as a whole, the MDA8 values for the TCEQ sites in Table B.2 were calculated as above. The CASTNet and IMPROVE data was already provided as appropriately averaged values. The scripts calc_TX_bkgrd_ozone.py and calc_TX_bkgrd_PM25.py were used to process the data.

B.2.3 Estimating Preliminary Background Values

The lowest of the daily MDA8 O_3 values in the background sites for each urban area were selected as our preliminary background estimates. In addition, the maximum and minimum MDA8 O_3 values for all urban sites in the area were also calculated.

Similarly, the lowest of the daily average $PM_{2.5}$ values in the background sites for each urban area were selected as our preliminary background estimates. In addition, the maximum and minimum daily average $PM_{2.5}$ values for all urban sites in the area were also calculated.

For the State of Texas as a whole, we calculated separate background values for O_3 from the TCEQ sites and the CASTNet sites. In both cases the minimum valid MDA8 value was used. The minimum valid IMPROVE $PM_{2.5}$ value was used as the $PM_{2.5}$ background estimate for Texas.

B.2.4 Linear Regressions Test and Outlier Analysis

We investigated the preliminary background estimates for each urban area by performing a linear regression of the preliminary background values (x) versus the maximum values (y) of MDA8 O_3 and daily average $PM_{2.5}$ using the R software package (using scripts bckgd_fit_o3.R and bckgd_fit_pm.R). For example, Figure B.1 shows a scatterplot of the background MDA8 O_3 value versus the maximum MDA8 O_3 value for the HGB area (the other fit figures are included in the original appendix attached to this deliverable). The solid black line is the linear fit, and the dotted and dashed black lines are the upper and lower 95% (or 2σ) confidence intervals, respectively. In this example, 89 of the 1834 valid data points (4.9%) have maximum MDA8 O_3

[6] All listed scripts will be supplied to TCEQ with the project's Final Report.

values that fall above the upper confidence internal of the linear fit, suggesting that these background estimates are lower than would be expected given the maximum values seen in the urban area. Table B.3 gives the number of such points for each urban area and pollutant. All such data points are identified in the csv files in a column called "high_flag", with a value of TRUE meaning that day was above the upper 95% confidence interval for that day. Given the skewed distribution of both the background and maximum MDA8 O_3 and daily average $PM_{2.5}$, very few points were identified below the lower confidence interval of the fit (one MDA8 O_3 value for DFW, six $PM_{2.5}$ values for ARR, and three $PM_{2.5}$ values for HGB) and so these points are not flagged in the csv files.

In addition, for some days only one monitoring site within the urban area had a valid MDA8 O_3 or daily average $PM_{2.5}$ value, so that the maximum and preliminary background estimates were identical. These data points are identified in the csv files in a column called "eq_flag", with a value of TRUE meaning that day only had a single site with valid data, and so the maximum and background estimates are equal (see also Section 3). For example, this is true of all background $PM_{2.5}$ estimates for the Tyler-Longview-Marshall area (TLM), as there was only a single site with valid $PM_{2.5}$ data (see Table B.3). ***While we have included these data points in our background estimates for completeness, we strongly recommend that users be careful about including these points in their analyses, as they may bias the results of, for example, the average difference between the maximum and background values.***

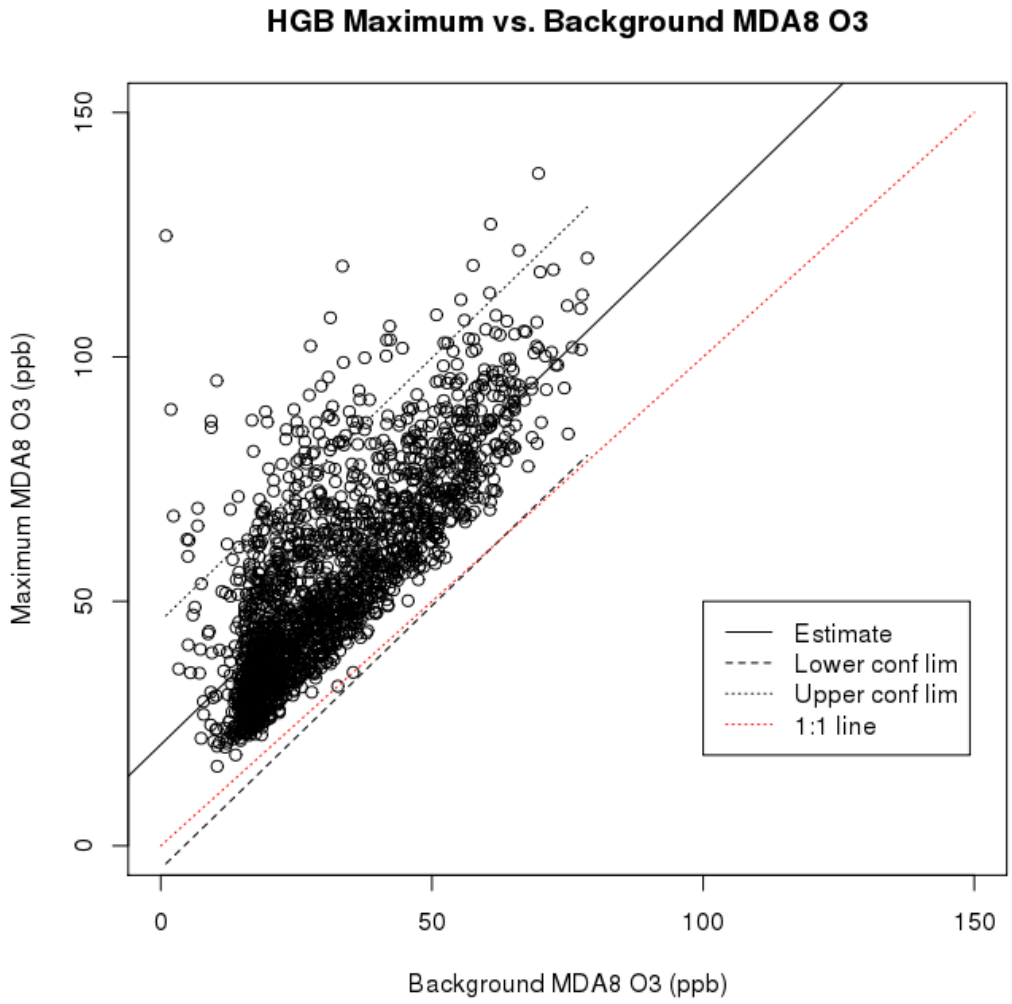


Figure B.1. Maximum versus background MDA8 O_3 values for the HGB area.

Table B.3. Number of background points quality flagged for each urban area and pollutant.

Urban Area	MDA8 O_3				Daily Average $PM_{2.5}$			
	# high	# eq	# final	# replaced	# high	# eq	# final	# replaced
DFW	78	0	46	0	118	21	68	0
HGB	89	2	61	43	149	11	101	75
SA	67	117	37	0	129	490	0	0
ARR	72	217	20	0	172	100	0	0
BPA	78	14	65	49	112	226	0	0
TLM	93	30	0	0	NA	3509	NA	NA

Similar to Berlin et al. (2013), we performed further analysis of the points that were above the 95% confidence interval of the fit (e.g., where high_flag = TRUE). First, we identified the subset of those points where (a) where high_flag = TRUE AND (b) at least one other background site in the urban area had a valid MDA8 O_3 or daily average $PM_{2.5}$ value for that day AND (c) the valid values at the other background sites were all more than 10% larger than the preliminary background estimate. Note that the later two criteria have to be true for replacing the preliminary background estimate with a value from a different background site to make a significant impact on any subsequent analysis. Data points that met all three criteria are flagged in the csv files in a column called "final_flag", with a value of TRUE meaning that the above criteria were satisfied. The number of points with final_flag = TRUE for each urban area is shown in Table B.3. For these points, we have included the AQS site number and the MDA8 O_3 or daily average $PM_{2.5}$ value for the background site with the second largest value in the csv files as an alternate background estimate.

However, we only replaced the preliminary background value if:

4. The final_flag = TRUE
5. The estimate was for the HGB or BPA areas, as these areas near the Gulf of Mexico could plausibly have times when the gulf/lake breeze front affects some of the outlying background sites, but does not affect the urban area as a whole.
6. The preliminary background site was between the city and the Gulf of Mexico (or the city and Sabine Lake). These sites are given in Table B.4.

Table B.4. Background sites that were replaced if final_flag = TRUE

Urban Area	AQS Site Numbers of Background Sites that could be replaced
HGB	480391016, 480390618, 481671034, 480390619
BPA	482450628, 482450101

The total number of background sites where data was replaced with the second highest value is given in Table B.3. These replacements do not significantly impact the statistics of the background estimates.

For the State of Texas as a whole, we did not perform a similar sort of analysis, as there is difficulty in deciding what is the appropriate maximum value to use for the state as a whole. The preliminary background estimates are thus identical to the final background estimates.

B.3 File Descriptions

The data contained in the csv files included in the deliverable are described below. All files are in comma-separated-value (csv) format unless otherwise stated.

B.3.1 Urban Area Ozone Files (file name = *_flagged_O3_v3.csv, six files in total)

Column Descriptions:

1. **Date:** In YYYYMMDD format. Note only dates in the ozone season (May-October) will have valid values.

2. **AQS_Code_max:** AQS site number of the site in the urban area with the maximum MDA8 ozone.
3. **O3_max..ppbv.:** Maximum of the valid MDA8 ozone (ppbv) values for all sites in the urban area.
4. **AQS_Code_min_max:** AQS site number of the site with the minimum valid MDA8 ozone (ppbv) for ***all*** sites in an urban area. Note this may not be equal to the background estimate (Columns 6 and 7), as that is the minimum for the ***background*** sites only.
5. **O3_min_max..ppbv.:** Minimum of the valid MDA8 ozone (ppbv) values for ***all*** sites in an urban area. Note this may not be equal to the background estimate (Columns 6 and 7), as that is the minimum for the ***background*** sites only.
6. **AQS_Code_min:** AQS site number of the preliminary background estimate.
7. **O3_min_bkgrd..ppbv.**: The preliminary background estimate, calculated as the minimum valid MDA8 ozone (ppbv) for the background sites in an urban area.
8. **high_flag:** TRUE if this day was above the 95% confidence interval for a linear fit of the preliminary background MDA8 ozone value (x, Column 7) against the maximum MDA8 ozone value (y, Column 3). See Section B.2.4.
9. **eq_flag:** TRUE if this day only had one valid MDA8 ozone value for the urban area, and so the preliminary background MDA8 ozone value (x, Column 7) and the maximum MDA8 ozone value (y, Column 3) are equal. ***We strongly recommend that users be careful about including the points flagged as TRUE in their analysis, as they may bias the results of, for example, the average difference between the maximum and background values.*** See Section B.2.4.
10. **final_flag:** TRUE if (a) high_flag (Column 8) is TRUE, AND (b) at least one other background site in the urban area had a valid MDA8 ozone value for that day, AND (c) the valid values at the other background sites were all more than 10% larger than the preliminary background estimate. See Section B.2.4.
11. **X2nd.Highest.MDA8.AQS.Code:** If final_flag = TRUE, this contains the AQS site number of the background site with the second lowest MDA8 value.
12. **X2nd.Highest.MDA8..ppbv.:** If final_flag = TRUE, this contains the second lowest MDA8 value of the background sites.
13. **Final.MDA8.Background.AQS.Code:** AQS site number of the final MDA8 background estimate (ppbv), with some HGB and BPA values replaced as described in Section 2.4.
14. **Final.MDA8.Background..ppbv.:** final MDA8 background estimate (ppbv), with some HGB and BPA values replaced as described in B.2.4.

B.3.2 Urban Area $PM_{2.5}$ Files (file name = *_flagged_PM_v2.csv, six files in total)

Column Descriptions:

1. **Date:** In YYYYMMDD format.
2. **AQS_Code_max:** AQS site number of the site in the urban area with the maximum daily average $PM_{2.5}$ value.
3. **PM2.5_max..ug.m.3.):** Maximum of the valid daily average $PM_{2.5}$ values ($\mu g\ m^{-3}$) for all sites in the urban area.

4. **AQS_Code_min_max:** AQS site number of the site with the minimum valid daily average $PM_{2.5}$ values ($\mu g\ m^{-3}$) for ***all*** sites in an urban area. Note this may not be equal to the background estimate (Columns 6 and 7), as that is the minimum for the ***background*** sites only.
5. **PM2.5_min_max..ug.m.3:** Minimum of the valid daily average $PM_{2.5}$ values ($\mu g\ m^{-3}$) for ***all*** sites in an urban area. Note this may not be equal to the background estimate (Columns 6 and 7), as that is the minimum for the ***background*** sites only
6. **AQS_Code_min:** AQS site number of the preliminary background estimate.
7. **PM2.5_min_bkgrd..ug.m.3.:** The preliminary background estimate, calculated as the minimum valid daily average $PM_{2.5}$ values ($\mu g\ m^{-3}$) for the background sites in an urban area.
8. **high_flag:** TRUE if this day was above the $95^{\%}$ confidence interval for a linear fit of the preliminary background MDA8 ozone value (x, Column 7) against the maximum MDA8 ozone value (y, Column 3). See Section 2.4.
9. **eq_flag:** TRUE if this day only had one valid MDA8 O3 value for the urban area, and so the preliminary background daily average $PM_{2.5}$ values ($\mu g\ m^{-3}$) (x, Column 7) and the maximum daily average $PM_{2.5}$ values ($\mu g\ m^{-3}$) (y, Column 3) are equal. ***We strongly recommend that users be careful about including the points flagged as TRUE in their analysis, as they may bias the results of, for example, the average difference between the maximum and background values.*** See B.2.4.
10. **final_flag:** TRUE if (a) high_flag (Column 8) is TRUE, AND (b) at least one other background site in the urban area had a valid daily average $PM_{2.5}$ value for that day, AND (c) the valid values at the other background sites were all more than 10% larger than the preliminary background estimate.
11. **X2nd.Highest.PM2.5.AQS.Code:** If final_flag = TRUE, this contains the AQS site number of the background site with the second lowest daily average $PM_{2.5}$ value.
12. **X2nd.Highest.PM2.5..ug.m.3.:** If final_flag = TRUE, this contains the AQS site number of the background site with the second lowest daily average $PM_{2.5}$ value.
15. **Final.PM2.5.Background.AQS.Code:** AQS site number of the final daily average $PM_{2.5}$ background estimate ($\mu g\ m^{-3}$), with some HGB and BPA values replaced as described in B.2.4.
16. **Final.PM2.5.Background..ug.m.3.:** Final daily average $PM_{2.5}$ background estimate ($\mu g\ m^{-3}$), with some HGB and BPA values replaced as described in B.2.4.

B.3.3 Texas O_3 Background (file name = TX_State_bkgrd_O3_calc.csv)

Column Descriptions:

1. **Date:** In YYYYMMDD format. Note only dates in the ozone season (May-October) will have valid values.
2. **AQS_Code_min:** AQS site number of the preliminary background estimate using TCEQ sites near the Texas border.
3. **O3_min(bkgrd):** The preliminary background estimate, calculated as the minimum valid MDA8 ozone (ppbv) for the TCEQ sites near the Texas border.
4. **CASTNet ID:** CASTNet ID code for the CASTNet site near the Texas border with the lowest MDA8 O_3 value for the day.

5. **CASTNet O3_min (bkgrd):** The CASTNet-based background estimate, calculated as the minimum valid MDA8 ozone (ppbv) for the CASTNet sites near the Texas border.

B.3.4 Texas $PM_{2.5}$ Background (file name = TX_State_PM_calc.csv)

Column Descriptions:

1. **Date:** In YYYYMMDD format. Note that as these estimates are based on IMPROVE network data, data is only available one out of every 3 days.
2. **IMPROVE ID:** IMPROVE ID code of the IMPROVE site near the Texas border with the lowest daily average $PM_{2.5}$ value.
3. **PM_min(bkgrd, ug/m^3):** The background estimate for the State of Texas, calculated as the minimum valid daily average $PM_{2.5}$ values ($\mu g\ m^{-3}$) for the background sites in an urban area.

B.4 Quality Assurance Steps

In addition to the analyses described in B.2.4, other quality assurance checks were made. First, all scripts used in this project were independently inspected to ensure they were calculating properly, and any errors noted in early versions were fixed. Second, the statistics of the background and maximum values for each urban area were investigated to ensure that they were reasonable and did not change unexpectedly between file versions.

Appendix C: File Descriptions in Final Deliverable Package

All associated data and scripts for this project are contained in the deliverable packages, which can be downloaded from the AER ftp server at:

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

ftp://ftp.aer.com/pub/malvarad/TCEQ/p1952_deliverable_3_1_R1_1.tar.gz

ftp://ftp.aer.com/pub/malvarad/TCEQ/p1952_deliverable_5_2_R2_0.tar.gz
```

The files contained in the packages for Deliverables 2.2 and 3.1 are documented in Sections A.2 and B.3, respectively. Here we discuss the additional files included in the final Deliverable 5.2.

C.1 *./P1952_trend_plots.xlsx*

This is a Microsoft Excel file (made using Microsoft Excel for Mac 2011 v14.5.3) that was used to produce the meteorologically adjusted annual averages and linear tends discussed in Section 2.4. The spreadsheet contains the data used to create the plots as well as the plots themselves.

C.2 Subdirectory *./MAPTYPE/*

This subdirectory contains the files used to perform the synoptic map type analyses and logistic regressions discussed in Section 4. The individual files are described below.

C.2.1 Map Type Files

C.2.1.1 ./MAPTYPE/narr_maptype_2005_2014_850_70_5types.dat

This is an ASCII text file that contains two columns, the date (MMDDYYYY, with no zeros as spacers) and the determined synoptic type for that date. The types are described in Section 4.1.1.

C.2.1.2 ./MAPTYPE/tceq_map_type.xlsx

This is a Microsoft Excel file (made using Microsoft Excel for Mac 2011 v14.5.3) that was used to convert the data in *./MAPTYPE/narr_maptype_2005_2014_850_70_5types.dat* into a comma-separated-value (CSV) text file with a date column that matches that in the CSV input files for the GAM fitting described in Section A.2.4.2.

C.2.1.3 ./MAPTYPE/tceq_map_type.csv

A CSV file produced from *./MAPTYPE/tceq_map_type.csv* that is used by ./MAPTYPE/syn_type_boxplot.R to merge the synoptic type data with the CSV input files for the GAM fitting described in Section A.2.4.2. The columns are Month, Day, Year, Syn.Type (Synoptic Type), and Date (in YYYYMMDD format).

C.2.2 R Scripts

C.2.2.1 ./MAPTYPE/syn_type boxplot.R

This script reads in *./MAPTYPE/tceq_map_type.csv* and the CSV input files for the GAM fitting described in Section A.2.4.2 and then produces box plots of how total and background O_3 and $PM_{2.5}$ vary between the synoptic types. It calculates the mean and standard deviation of O_3 and $PM_{2.5}$ for each type, as well as the percentage of days in each type above a fixed threshold (70 ppb for total MDA8 O_3, 55 ppb for background MDA8 O_3, 17 ug/m^3 for total daily $PM_{2.5}$ and

13 ug/m^3 for background daily $PM_{2.5}$). It also fits the log of the concentrations to a linear model of the synoptic types.

Text outputs are written to the log file *./MAPTYPE/syn_type_boxplot.log*, while the box plots for each city are saved to the files *./MAPTYPE/syn_type_boxplot*.png*, where the * is the city name.

Finally, the code produces updated GAM data files *(./MAPTYPE/*_merged_GLM_all_type_exceed.csv*) with additional columns that identify the synoptic types and the days that had values above the fixed threshold. These files are used as input by the script *./MAPTYPE/logistic_regress.R*.

C.2.2.2 ./MAPTYPE/logistic_regress.R

This script reads in the data files produced by *./MAPTYPE/syn_type_boxplot.R (./MAPTYPE/*_merged_GLM_all_type_exceed.csv*) and performs a logistic regression to determine how the probability of O_3 and $PM_{2.5}$ exceeding certain thresholds varies with meteorology, as described in Section 4.2.1. The output files are stored in separate subdirectories for each city (e.g., *./MAPTYPE/Houston/*). These include:

- A text log file (e.g., *./MAPTYPE/Houston/logistic_regress_Houston.log*)
- Probability plots for each pollutant metric (e.g., *./MAPTYPE/Houston/log_regress_plot_Houston_o3.max.exceed.png* is the same as Figure 25)
- An R data file (e.g., *./MAPTYPE/Houston/.RData_logistic_gam_Houston*) that contains the fitted model objects for each pollutant metric.

C.2.3 Updated GAM data files *(./MAPTYPE/*_merged_GLM_all_type_exceed.csv)*

These are CSV files created by *./MAPTYPE/syn_type boxplot.R* and used as input by *./MAPTYPE/logistic_regress.R*. They are identical to the files described in Section A.2.4.2 except for the addition of the following columns

- Month, Day, Year, and Syn.Type, following the format described in Section C.2.1.3.
- o3.max.exceed, o3.bg.exceed, pm.max.exceed, pm.bg.exceed: These are logical arrays that describe if the given pollutant metric (total MDA8 O_3, background MDA8 O_3, total daily average $PM_{2.5}$, background daily average $PM_{2.5}$, respectively) is equal to or greater than (TRUE) or less than (FALSE) the thresholds described in Section 4.1 (70 ppb for total MDA8 O_3, 55 ppb for background MDA8 O_3, 17 ug/m^3 for total daily $PM_{2.5}$ and 13 ug/m^3 for background daily $PM_{2.5}$).

C.2.4 R Output Files for Logistic Regression *(./MAPTYPE/*/RData_logistic_gam*)*

This is an R data file that contains the following R variables

- *mod*: A list containing the four GAM model objects made by the logistic regression of each of the logical arrays described in Section C.2.3
- *modeled.vars*: An array of the names of the four logical arrays used for the logistic regression. These are also the names of the four model objects in *mod*.
- City: The city name.

- Temps: The array of afternoon mean temperatures (^{o}C) used to produce the output probability plots.
- Winds: The array of daily average wind speeds (m/s) used to produce the output probability plots.
- Types: The array of synoptic types (see Section 4.1) used to produce the output probability plots.

C.3 Subdirectory *./PCA/SCRIPTS*

This directory contains the following python scripts used to perform the PCA analysis of O_3 and $PM_{2.5}$ in each urban area and use the results to calculate background estimates.

- */calc_PCA_bkgrd_ozone.py*: This script takes the raw TCEQ measurement data for all sites. It then calculates the MDA8 ozone for all sites and filters out those that have less than 75 % of the data for the ozone season during the 10-year time-span. It creates the .csv files ready to be spatially interpolated in interp_PCA_bkgrd_ozone.py
- */interp_PCA_bkgrd_ozone.py:* This file takes the MDA8 ozone file created in calc_PCA_bkgrd_ozone.py and spatially interpolates any missing datapoints by lat and lon. It creates the .csv files that are to be used for the PCA analysis in R pca_script.R
- */compare_bkgrdO3.py*: This script reads in the final PCA-predicted background O_3 values and plots the results compared to the original TCEQ-method predicted background O_3 and calculates the correlation statistics.
- */calc_PCA_PM2.5_bkgrd.py:* This script takes the raw TCEQ measurement data for all sites. It then calculates the daily average $PM_{2.5}$ for all sites and filters out those that have less than 75 % of the data for the entire year during the 10 year time-span. It creates the .csv files ready to be spatially interpolated in interp_PCA_bkgrd_PM25.py
- */interp_PCA_bkgrd_PM25.py*: This file takes the $PM_{2.5}$ file created in calc_PCA_bkgrd_ozone.py and spatially interpolates any missing data points by lat and lon. It creates the .csv files that are to be used for the PCA analysis in R: pca_script.R
- */compare_bkgrdPM.py:* This script reads in the final PCA-predicted background $PM_{2.5}$ values and plots the results compared to the original TCEQ-method predicted background $PM_{2.5}$ and calculates the correlation statistics.
- */pca_script.R:* This script follows the Eigenvector calculation to do a PCA analysis on the data sets created in interp_PCA_bkgrd_ozone.py and interp_PCA_bkgrd_PM.py it then outputs the completed background calculation to the txt files to be used in compare_bkgrdO3.py and compare_bkgrdPM.py.
- */plot_TCEQ_vs_PCA_mean.py:* This script reads in the final PCA and TCEQ determined background cases. It calculates the least squares of the yearly trends for each year and each of the four Group 1 urban areas.

C.4 Subdirectory *./PCA/FILES/O3*:

- **_Lat_calc.csv, *_Lon_calc.csv:* These files contain the latitude and longitude of each of the sites selected for the final interpolation and PCA analysis for each city.
- **MDA_O3_calc.csv*: These files contain the pre-interpolated MDA8 ozone values for the sites in that urban area that passed the criteria of having enough data points for the entire analysis time span.

- **_interp_O3.csv*: These files contain the post-interpolated MDA8 values for the sites in that urban area that passed the criteria of having enough data points for the entire analysis time span. Interpolation was done for missing data through either a cubic interpolation or nearest-neighbor interpolation, depend if the Latitude/Longitude of that station was not or was within the cluster of sites that did have data for that day.
- **_pca_derived_bkgrdO3_PC1.txt*: These files contain the post-processed, PCA calculated background ozone concentrations. Only one column of data is present corresponding to the one background estimated PCA value. The row corresponds to the date (first column) in the *_interp_O3.csv files above.

C.5 Subdirectory *./PCA/FILES/PM2.5/*

- **_Lat_calc.csv, *_Lon_calc.csv:* These files contain the latitude and longitude of each of the sites selected for the final interpolation and PCA analysis for each city.
- **_PM25_calc.csv*: These files contain the pre-interpolated daily average $PM_{2.5}$ values for the sites in that urban area that passed the criteria of having enough data points for the entire analysis time span.
- **_interp_PM25.csv*: These files contain the post-interpolated daily average $PM_{2.5}$ values for the sites in that urban area that passed the criteria of having enough data points for the entire analysis time span. Interpolation was done for missing data through either a cubic interpolation or nearest-neighbor interpolation, depend if the Latitude/Longitude of that station was not or was within the cluster of sites that did have data for that day.
- **_pca_derived_bkgrdPM_PC1.txt*: These files contain the post-processed, PCA calculated background ozone concentrations. Only one column of data is present corresponding to the one background estimated PCA value. The row corresponds to the date (first column) in the **_interp_PM25.csv* files above.

C.6 Subdirectory *./full_gam_fits*:

This directory contains the output files for the background O_3 and $PM_{2.5}$ GAM fits discussed in Section 2.3. The format of the files in this directory follows the format for the other GAM output files discussed in Section A.2.6, with the subdirectories labeled as *back_o3gam03_extended* and *back_pmgam03_extended*.

Appendix D: Logistic Regression Probability Plots for DFW, SA, and ARR

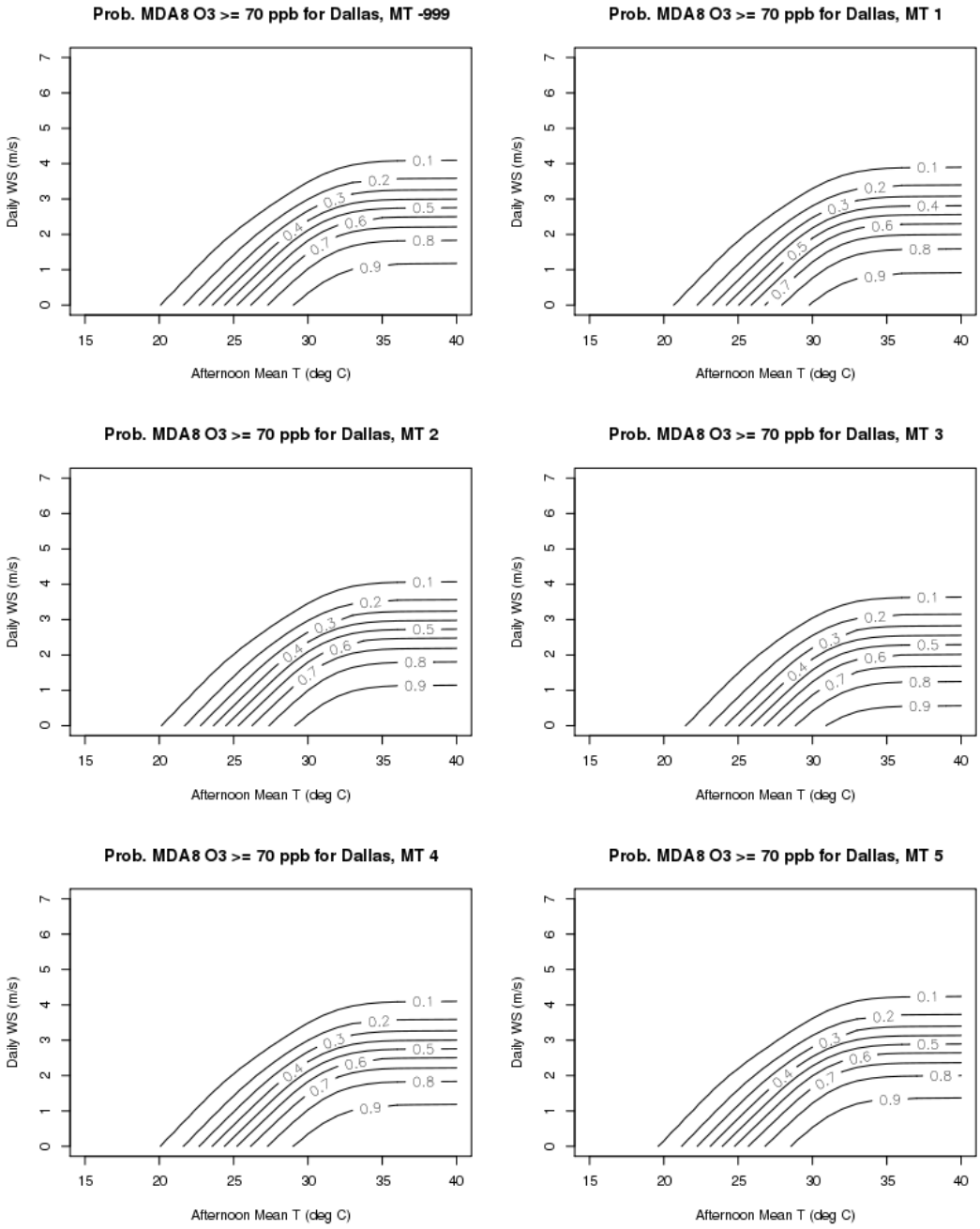


Figure D.1. Probability of the total MDA8 O_3 exceeding 70 ppbv for the Dallas/Fort Worth urban area as a function of afternoon mean temperature (^oC), daily wind speed (m/s), and synoptic type (as defined in Section 4.1).

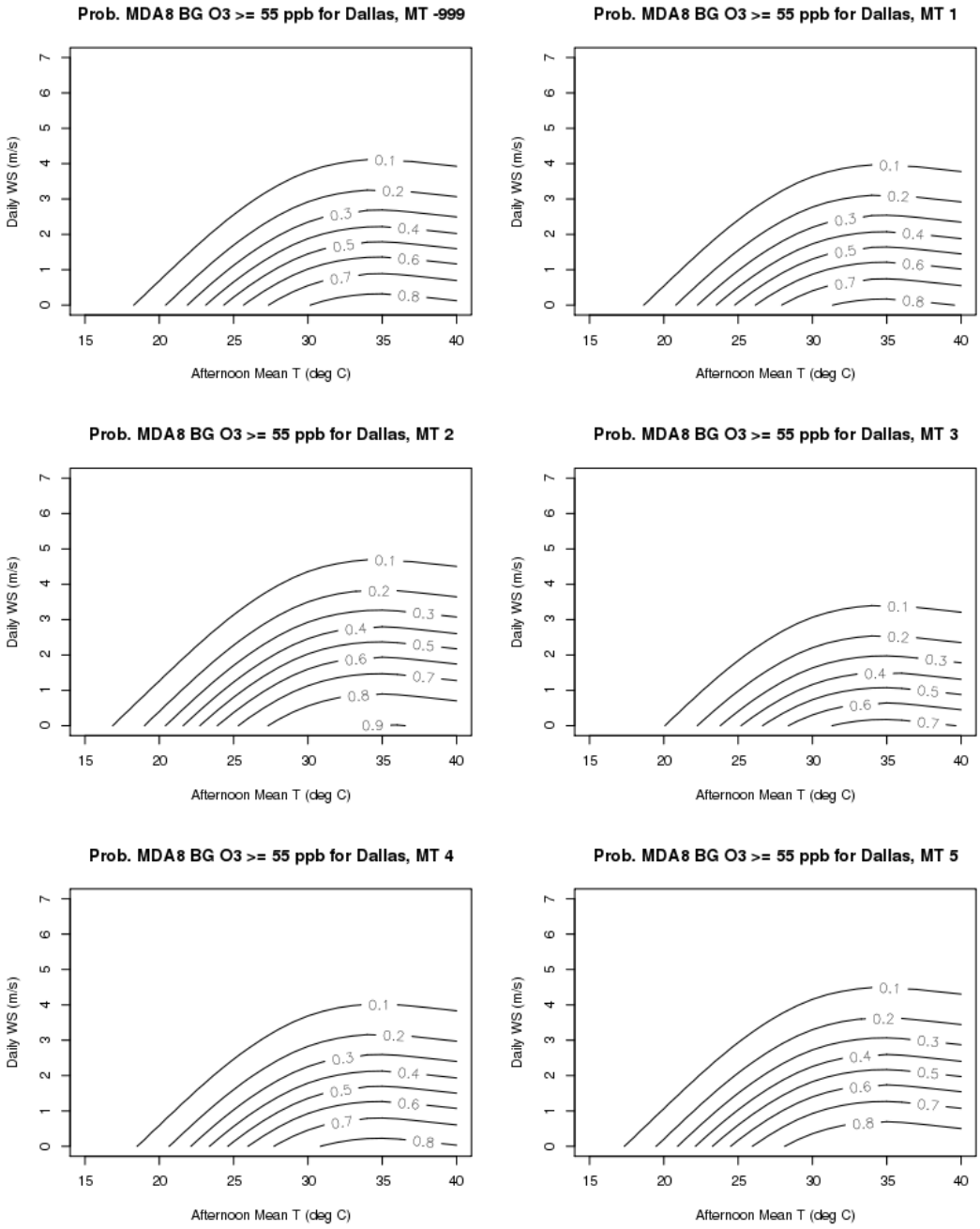


Figure D.2. As in Figure D.1, but for the probability of background MDA8 O_3 exceeding 55 ppbv.

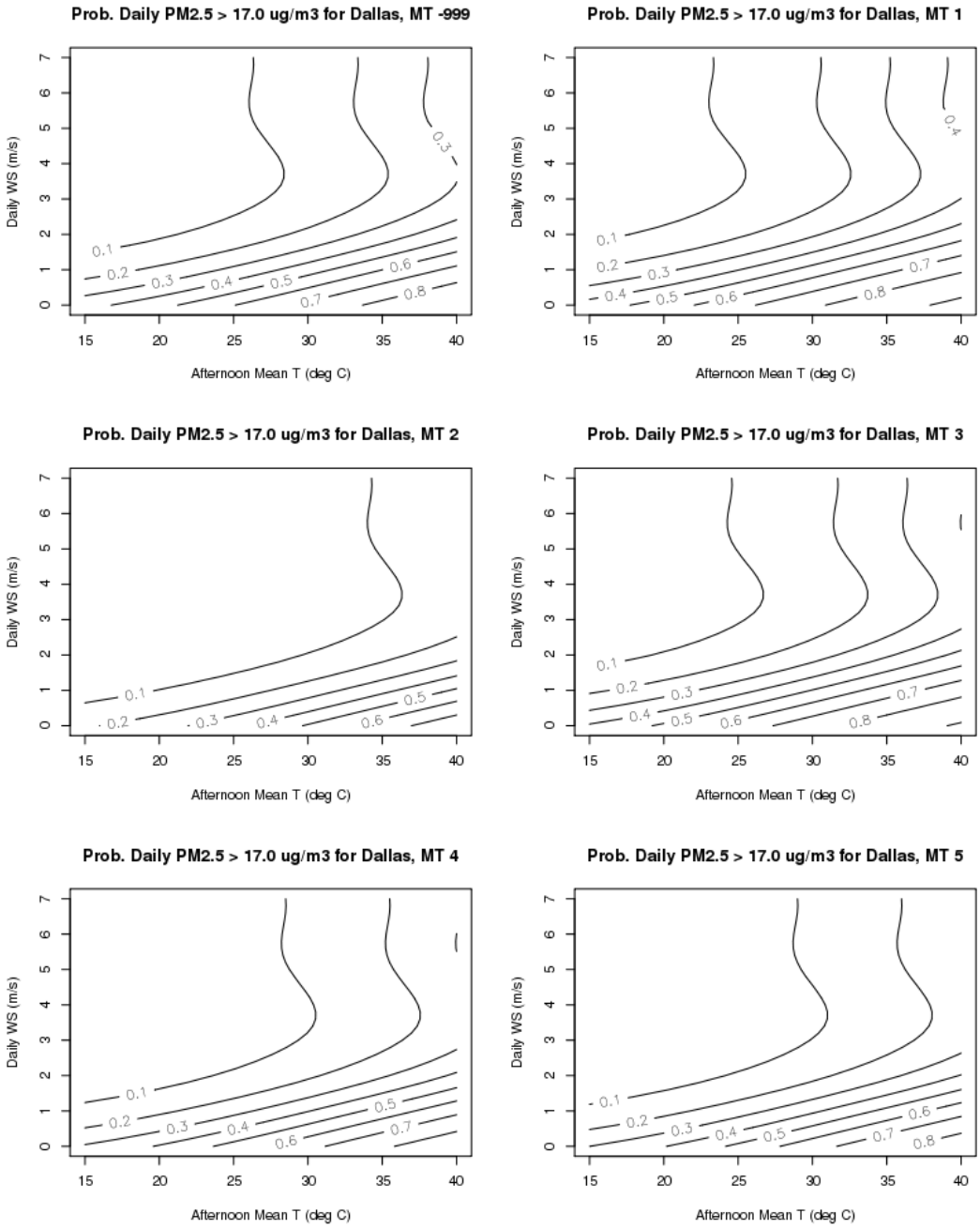


Figure D.3. As in Figure D.1, but for the probability of total daily average $PM_{2.5}$ exceeding 17 $\mu g/m^3$.

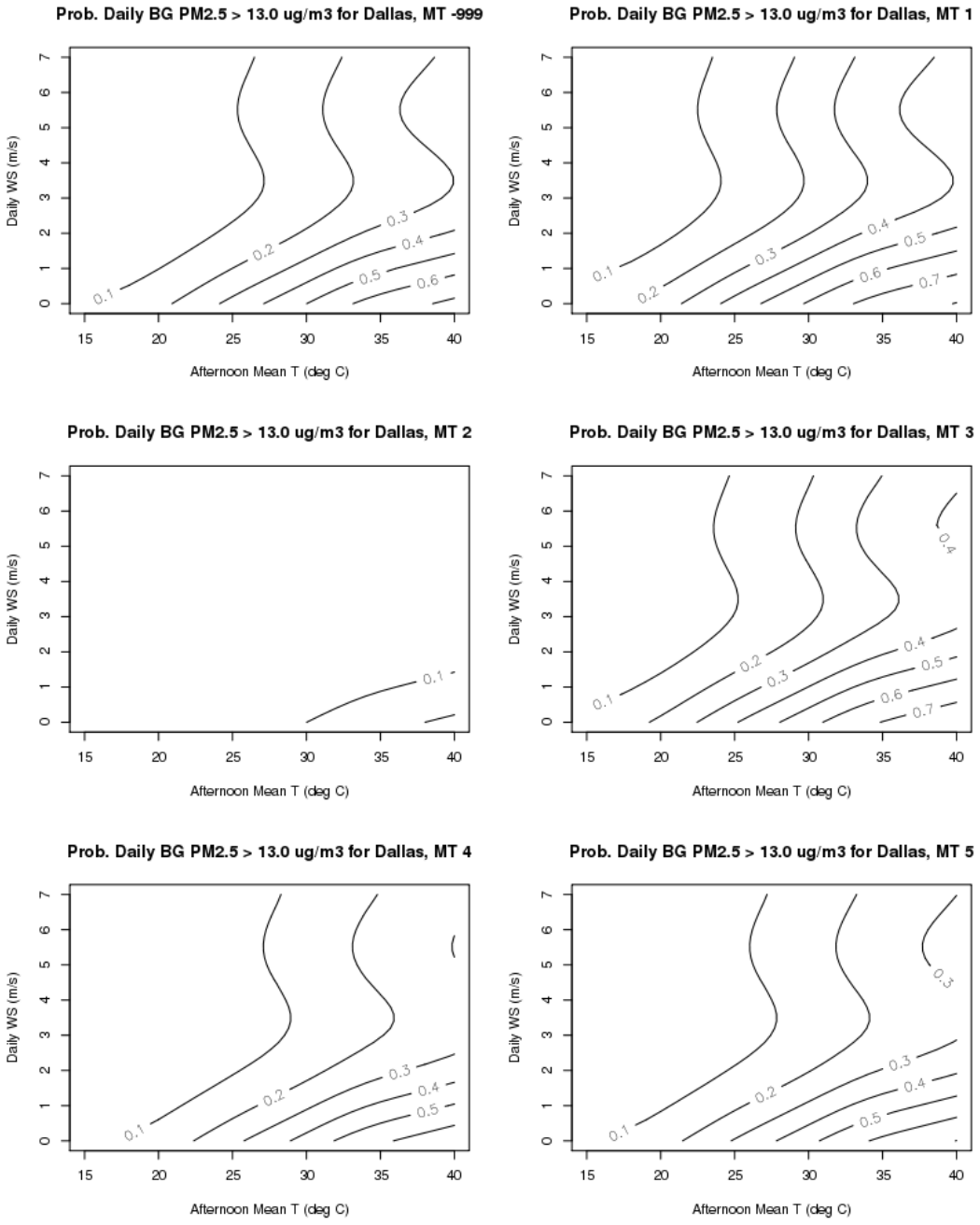


Figure D.4. As in Figure D.1, but for the probability of background daily average $PM_{2.5}$ exceeding 13 $\mu g/m^3$.

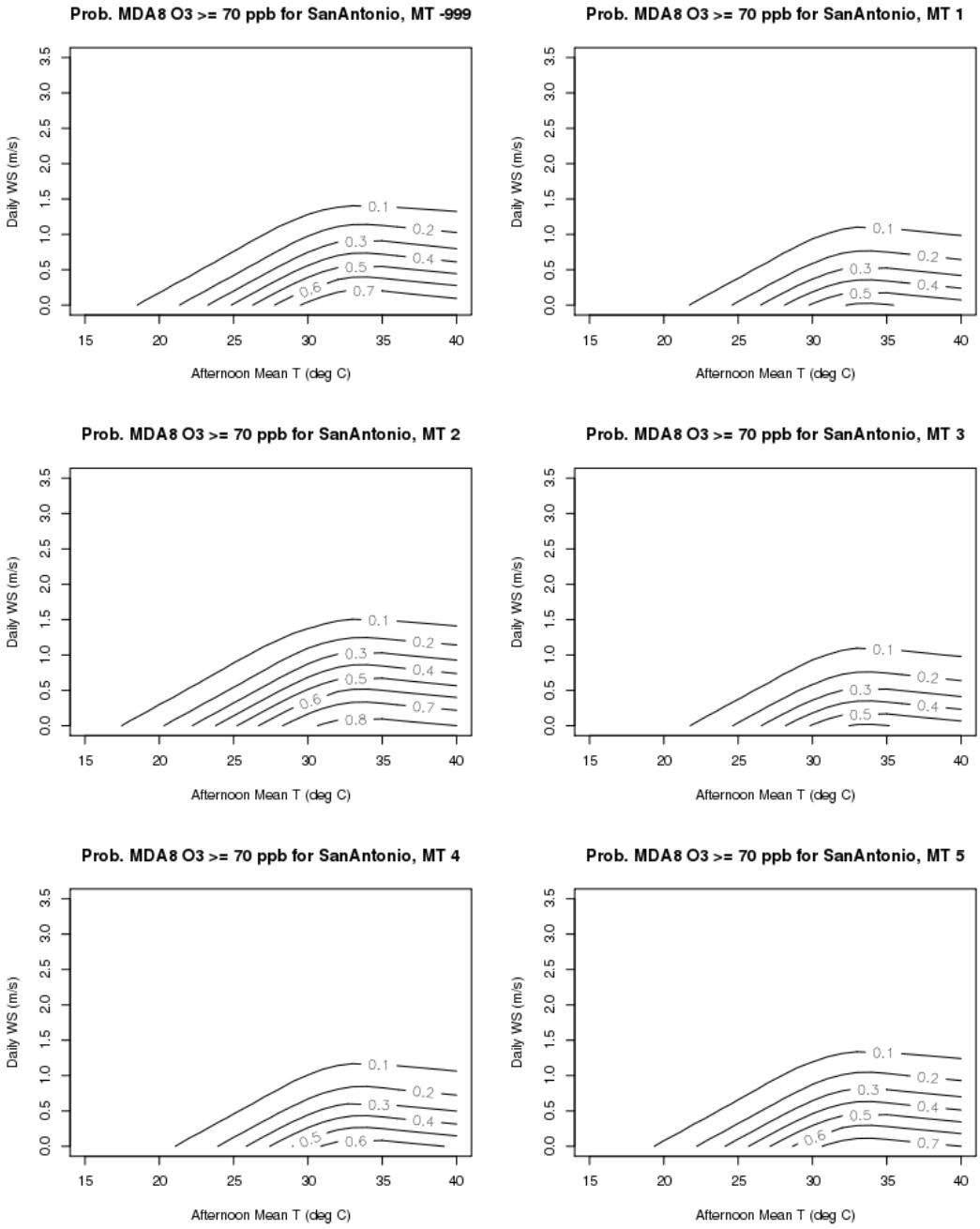


Figure D.5. As in Figure D.1 but for the San Antonio urban area.

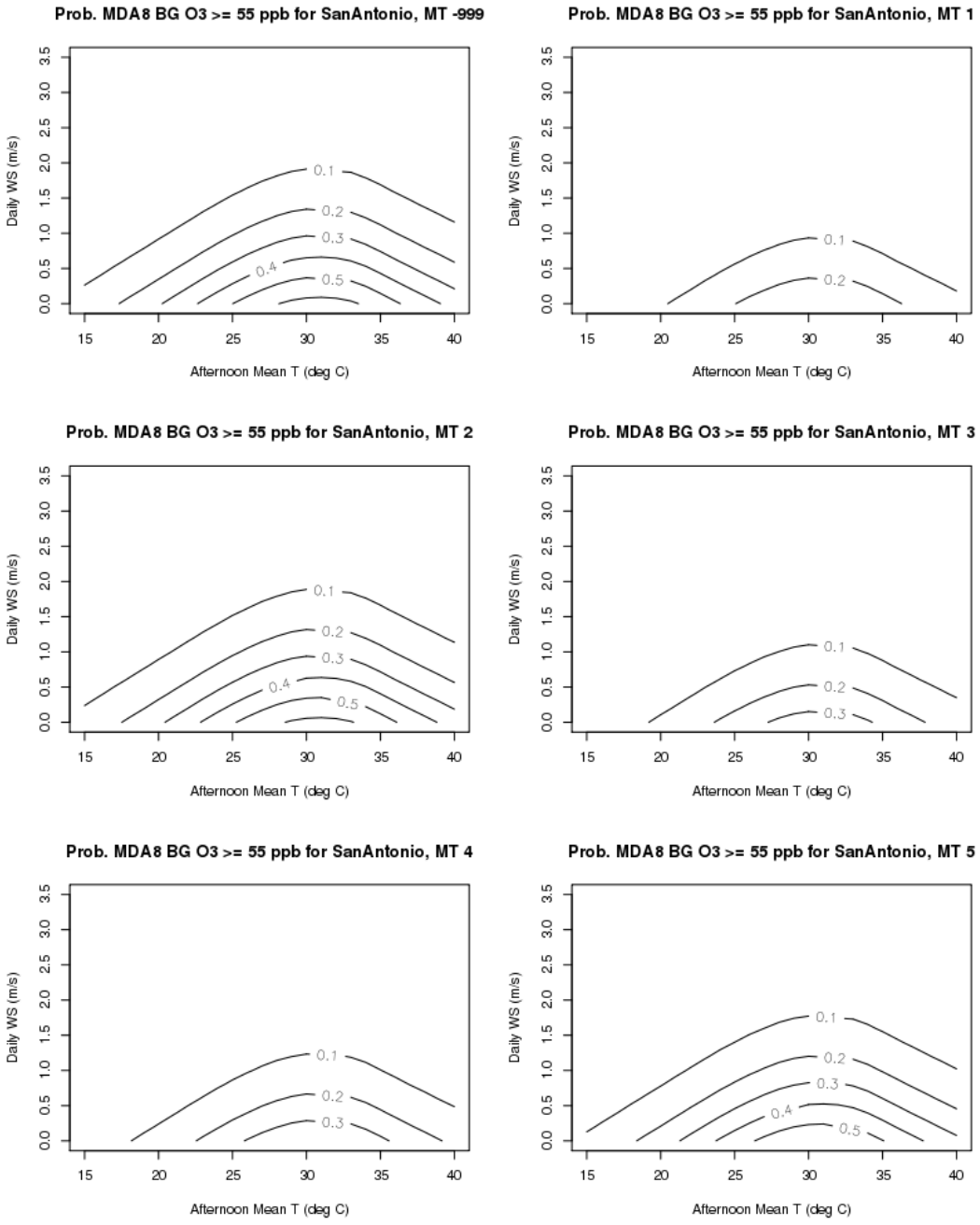


Figure D.6. As in Figure D.2 but for the San Antonio urban area.

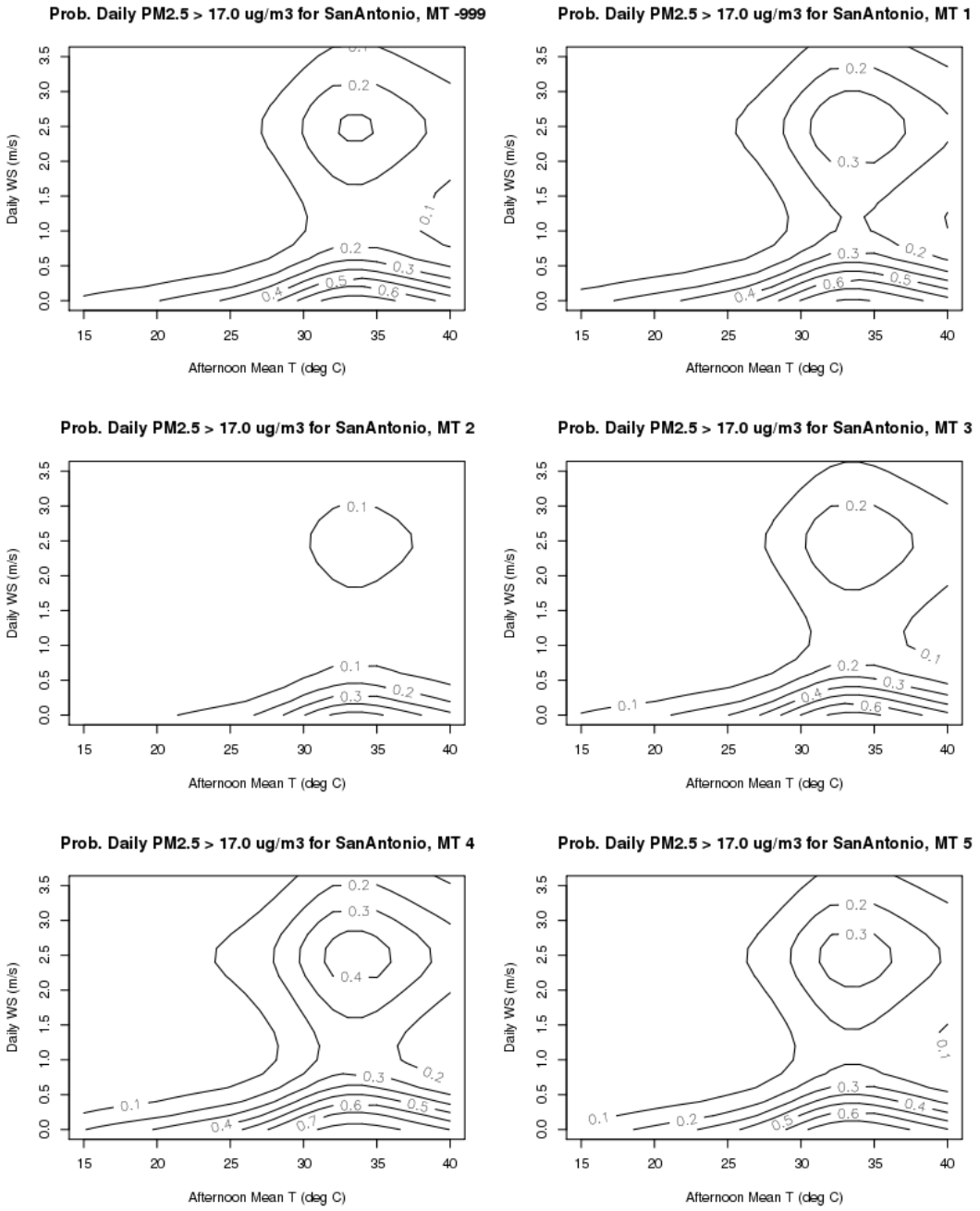


Figure D.7. As in Figure D.3 but for the San Antonio urban area.

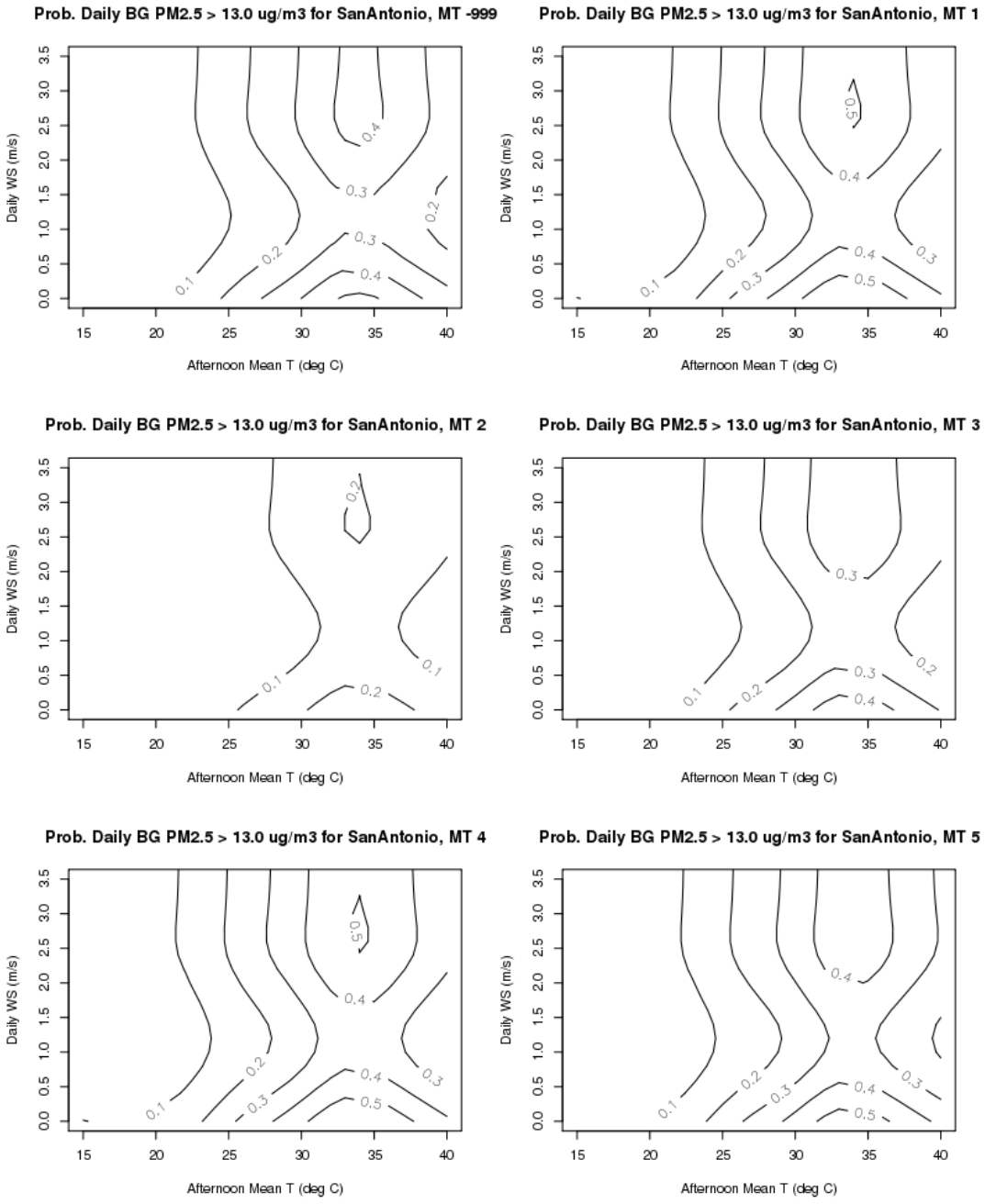


Figure D.8. As in Figure D.4 but for the San Antonio urban area.

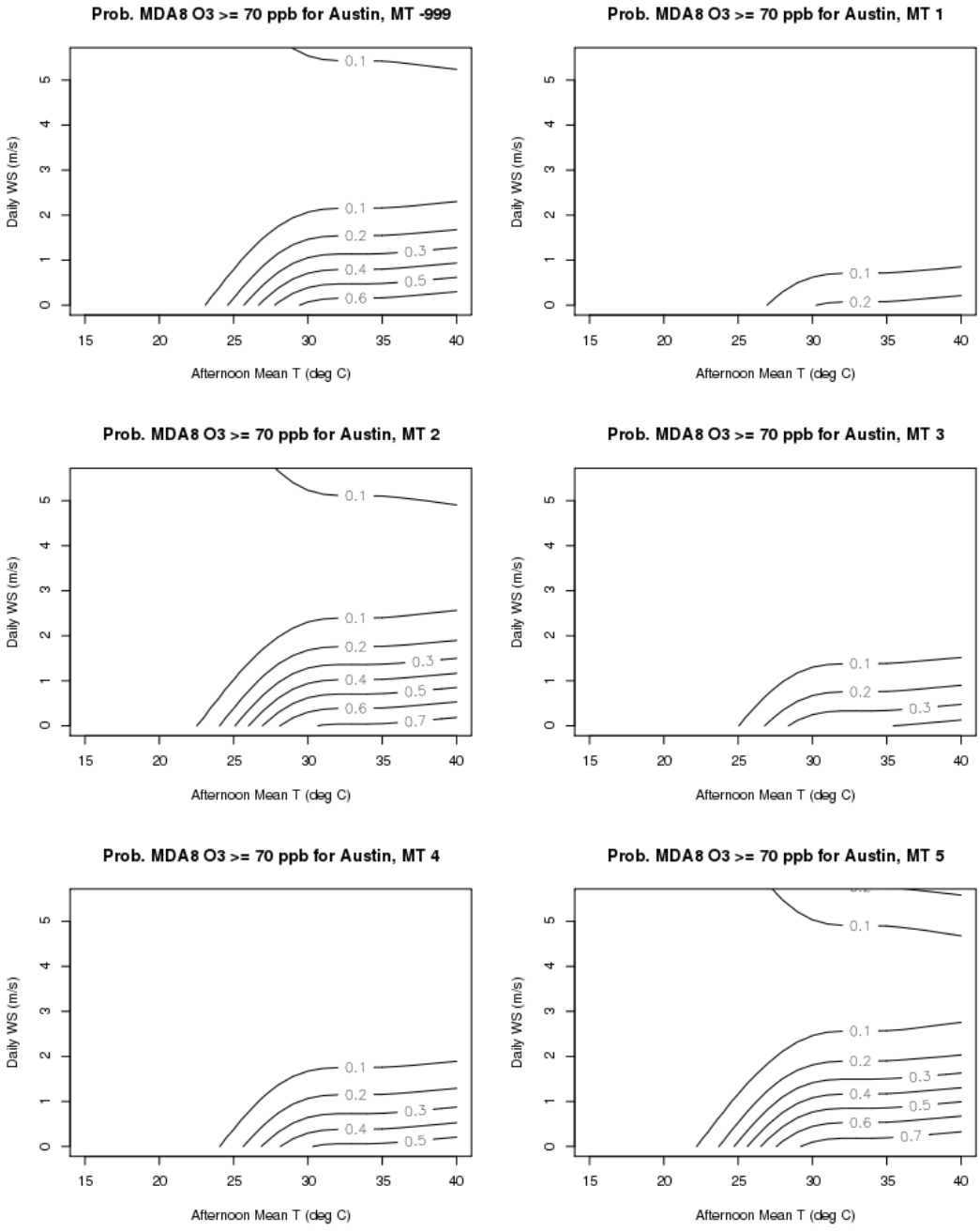


Figure D.9. As in Figure D.1 but for the Austin/Round Rock urban area.

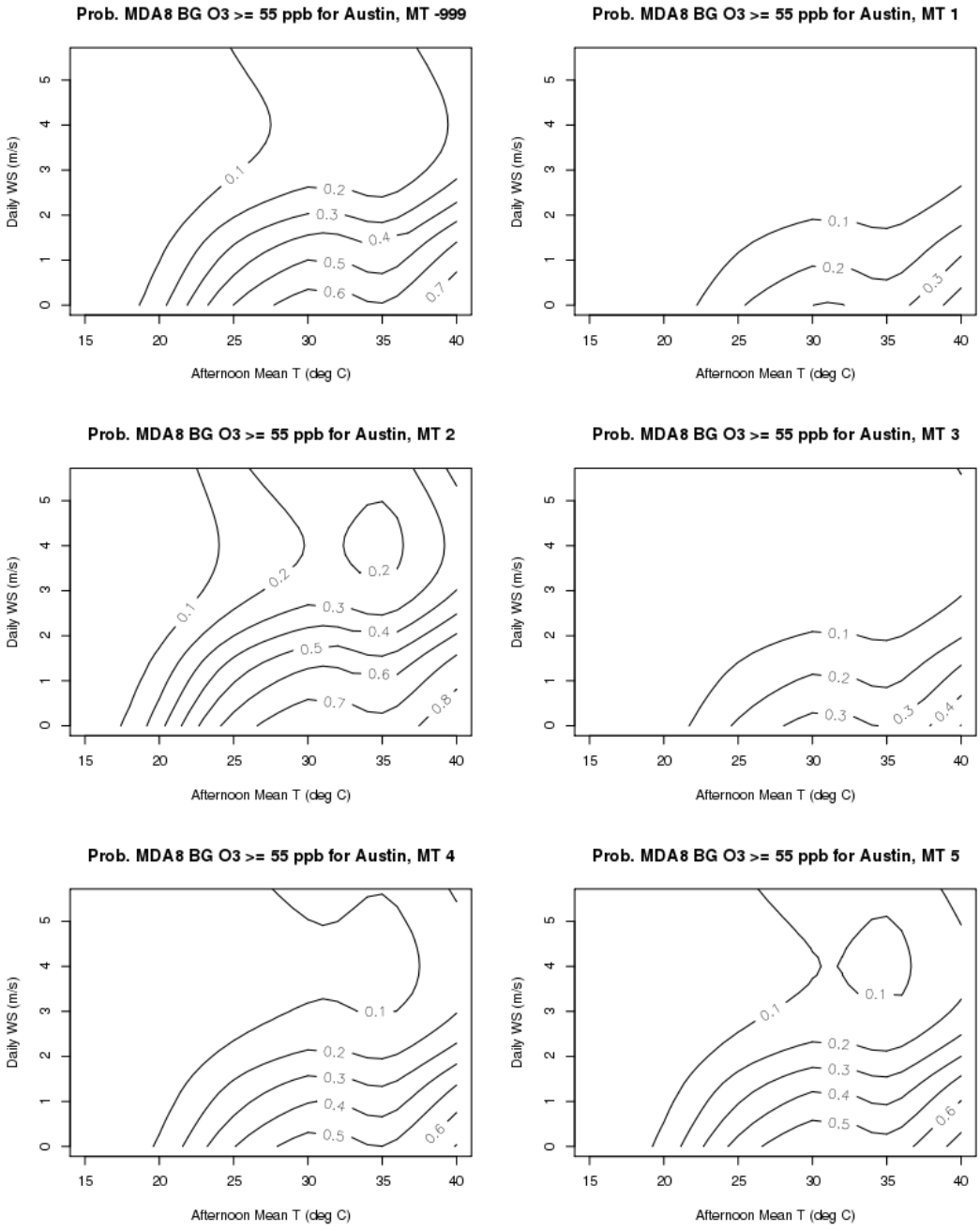


Figure D.10. As in Figure D.2 but for the Austin/Round Rock urban area.

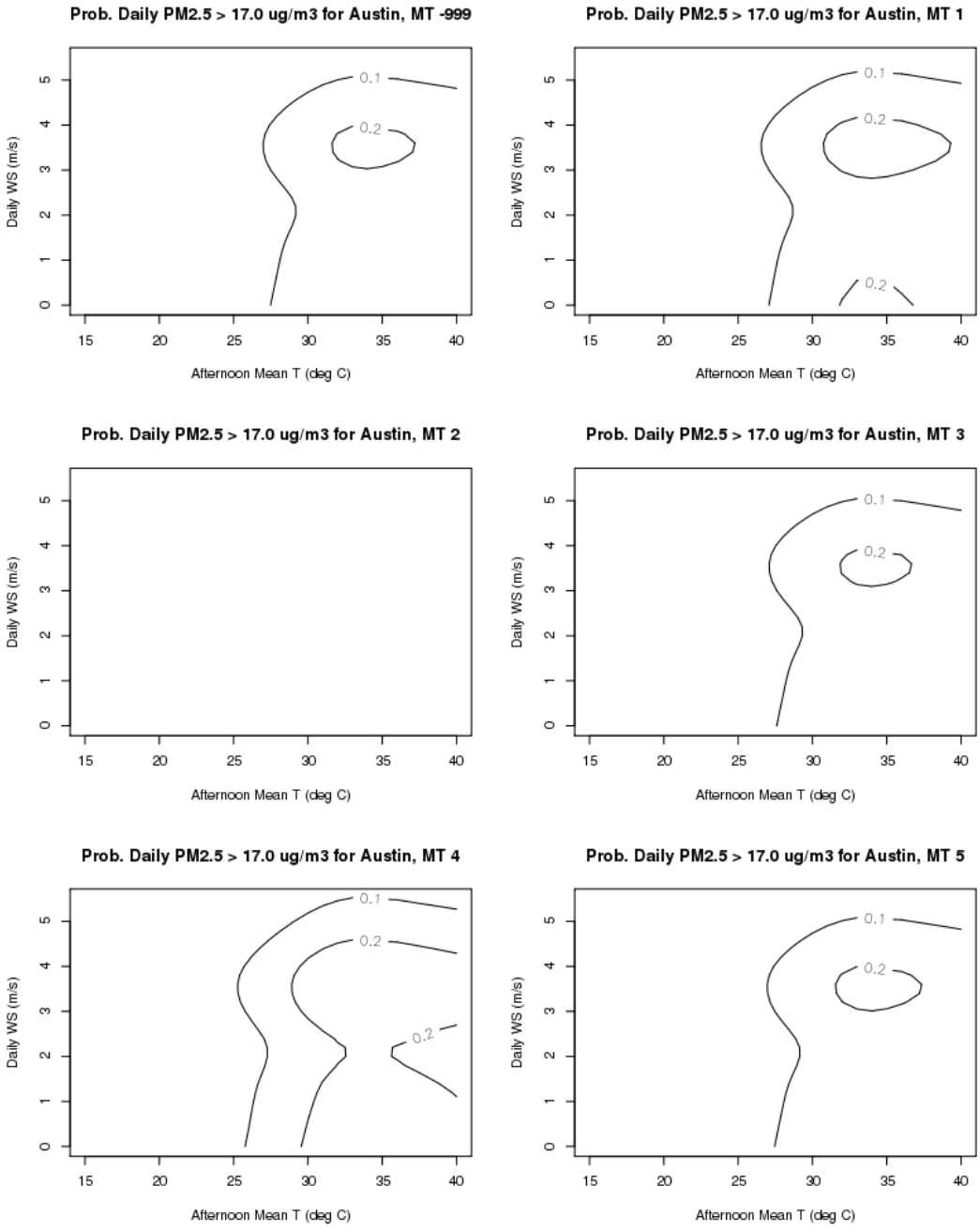


Figure D.11. As in Figure D.3 but for the Austin/Round Rock urban area.

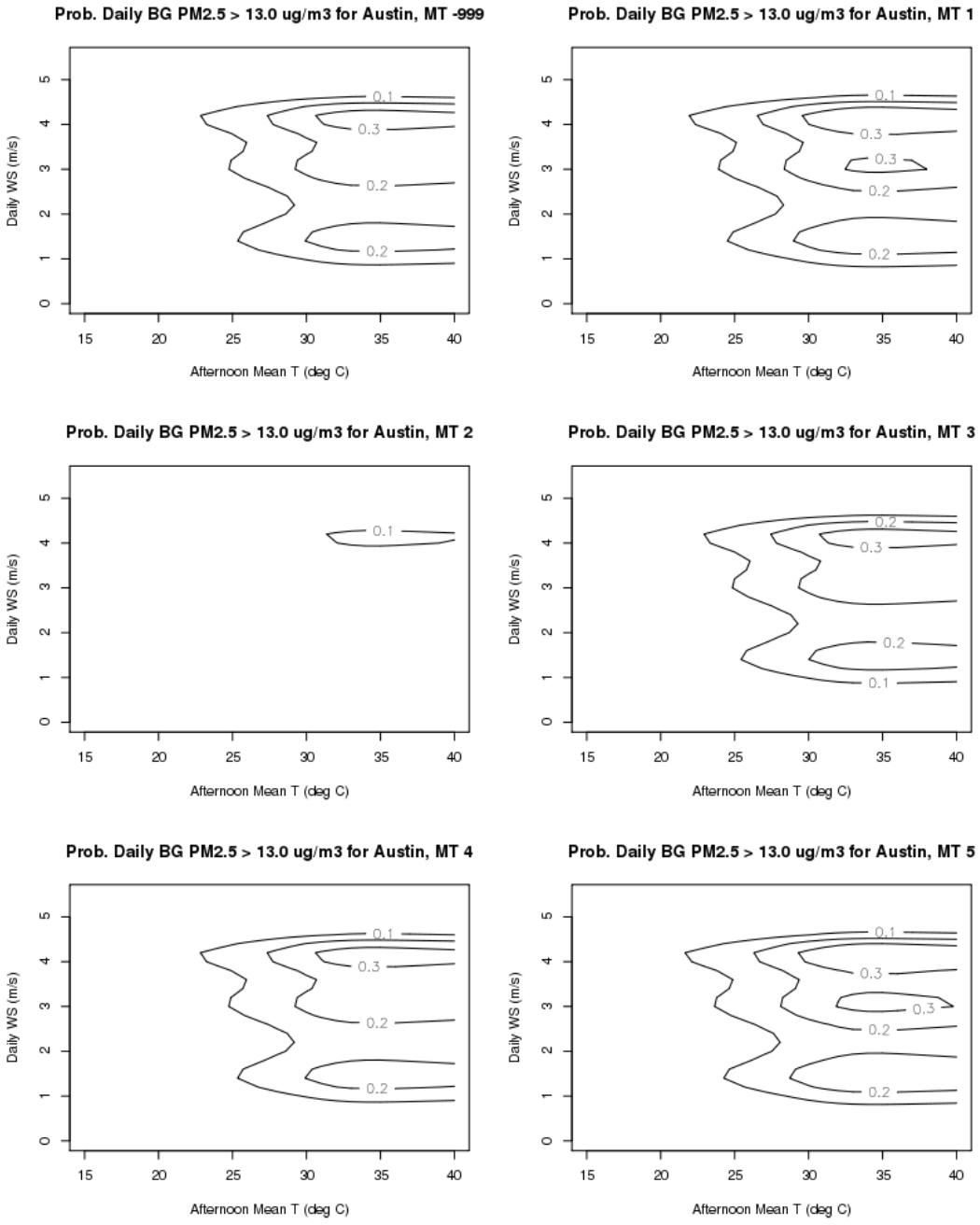


Figure D.12. As in Figure D.4 but for the Austin/Round Rock urban area.