The Pennsylvania State University
The Graduate School

ENHANCING THE VALUE OF AIR QUALITY FORECASTS IN THE MID-ATLANTIC REGION THROUGH USE OF ENSEMBLE STATISTICAL POST-PROCESSING

A Dissertation in
Meteorology
by
Gregory George Garner

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2013

The dissertation of Gregory George Garner was reviewed and approved* by the following:

Anne M. Thompson
Professor of Meteorology
Dissertation Adviser
Chair of Committee

William H. Brune
Professor of Meteorology
Head of the Department of Meteorology

George S. Young
Professor of Meteorology

Klaus Keller
Associate Professor of Geosciences

William F. Ryan
Research Assistant - Department of Meteorology
Special Member

*Signatures are on file in the Graduate School.

Abstract

Ozone pollution posses a significant threat to the health and well-being of the denizens in the mid-Atlantic region through degradation of the local air quality. Decision-makers rely on air quality forecasts to produce informed decisions that reduce the emission of and exposure to pollution. In these decision scenarios, air quality forecasts provide value of information. To increase the value of air quality forecasts in the mid-Atlantic region, an ensemble statistical post-processor (ESP) is designed. An analysis of the current value of information provided by air quality forecasts establishes a baseline to which the forecast using the ESP is compared. Air quality forecasts produced by the National Air Quality Forecast Capability (NAQFC), human air quality forecasters, and persistence are evaluated for predictive skill and economic value when used to inform decisions regarding pollutant emission and exposure. Surface ozone forecasts and observations are collected from 40 monitors representing eight forecast regions throughout Washington D.C., Virginia, and Maryland over the 2005 - 2009 ozone seasons (April - October). The value of the forecasts are quantified using a decision model based on costs to protect the public against a poor air quality event and the losses incurred if no protective measures are taken. The results indicate that the most skillful forecast method is not necessarily the most valuable forecast method. Air shed managers need to consider multiple forecast methods when deciding on multiple protective measures, because a single measure of forecast skill can often hide the user's sensitivity to forecast error for a specific decision.

Second, the next-generation numerical air quality model is assessed for value of information to determine a goal for the ensemble statistical post-processor. The NAQFC and an experimental version of the NAQFC (NAQFC-β) provided flight decision support during the July 2011 NASA DISCOVER-AQ field campaign around Baltimore, Maryland. Ozone forecasts from the NAQFC and NAQFC-β are compared to surface observations at six air quality monitoring stations in the DISCOVER-AQ domain. A bootstrap algorithm is used to test for significant bias and error in the forecasts from each model. The NAQFC-β tends to produce an average background ozone mixing ratio of at least 3.51 ppbv greater than the NAQFC throughout the domain at 95% significance. The difference between the two models is significant during the overnight and early morning hours likely due to the way the Carbon Bond 5 mechanism in the NAQFC-β handles reactive nitrogen recycling and organic peroxide species. The value of information each model provides is tested using a static cost-loss ratio model. By standard measures of forecast skill, the NAQFC generally outperforms the NAQFC-β; however, the NAQFC-β provides greater value of information. Five of the six sites exhibit an increase of value between 20% - 70% at low cost-loss ratios when using the NAQFC-β. The NAQFC generally produces equal or greater value for higher cost-loss ratio decisions.

Finally, the ESP is developed for the NAQFC to address the unique challenges of forecasting surface ozone in Baltimore, MD. Ozone and meteorological data are collected from the eight

monitors that constitute the Baltimore forecast region. These data are used to build the ESP using a moving-block bootstrap, regression tree models, and extreme-value theory. The ESP is evaluated using a 10-fold cross-validation to avoid evaluation with the same data used in the development process. Results indicate that the ESP is conditionally biased, likely due to slight overfitting while training the regression tree models. When viewed from the perspective of a decision-maker, the ESP provides a wealth of additional information previously not available through the NAQFC alone. The user is provided the freedom to tailor the forecast to the decision at hand by using decision-specific probability thresholds that define a forecast for an ozone exceedance. Taking advantage of the ESP, the user not only receives an increase in value over the NAQFC, but also receives value for costly decisions that the NAQFC couldn't provide alone.

Table of Contents

List of Figures

List of Tables

Preface

This dissertation is compiled from two published journal articles and one manuscript that has been submitted for publication.

Chapter 2 Garner, G.G., Thompson, A.M., 2012. The value of air quality forecasting in the mid-atlantic region. Wea. Climate Soc. 4, 69-79, DOI:10.1175/WCAS-D-10-05010.1.

Chapter 3 Garner, G.G., Thompson, A.M., Lee, P., Martins, D.K., 2013. Evaluation of naqfc model performance in forecasting surface ozone during the 2011 discover-aq campaign. Journal of Atmospheric Chemistry , 1-19, DOI:10.1007/s10874-013-9251-z.

Chapter 4 Garner, G.G., Thompson, A.M., 2013. Ensemble statistical post-processing of the National Air Quality Forecast Capability: Enhancing ozone forecasts in Baltimore, Maryland. Atmospheric Envirionment. Submitted.

Gregory Garner is the lead-author on each of these articles. The co-author contributions on each of these articles entails provision of data and/or background information or proof-reading. All hypotheses, analyses, and conclusions drawn from these articles are those of the lead author.

This research was supported by a STAR fellowship (FP-91729901) to Gregory Garner awarded by the U.S. Environmental Protection Agency. It has not been formally reviewed by the EPA. The views expressed in this dissertation are solely those of Gregory Garner. The EPA does not endorse any products or commercial services mentioned in this dissertation.

Acknowledgments

I would like to acknowledge my adviser Anne Thompson for her guidance and insight during my dissertation work. I would also like to acknowledge my committee members William Brune, George Young, Klaus Keller, and William Ryan for their patience and expertise. Laura Warren from the Maryland Department of the Environment and Dan Salkovitz from the Virginia Department of Environmental Quality have provided a significant amount of feedback on the ESP and perspective on the responsibilities of air quality forecasters in the Mid-Atlantic. I would like to thank the current and former Gators for keeping me focused and motivated. Finally, I would like to thank my family and friends for bearing with me throughout the years.

This work was funded by grants provided to Penn State University from NSF (DRU Program Award 0729413), NASA (AQAST: NNX11AQ44G, DISCOVER-AQ: NNX10AR39G) and the EPA (STAR: FP-91729901). Additional funding was provided by the NASA Tropospheric Chemistry Program and Aura Validation.

Dedication

I would like to dedicate this work to my family and friends. Without their patience and support, I would not have been able to complete this life-long goal. Thank you!

Chapter 1

Introduction

The goal of this dissertation is to develop an air quality forecast tool that can be optimized for the accuracy of forecasts and value of information under various decision scenarios. This work has been broken down into three broad tasks. The first task is to assess the current state of air quality forecasting with respect to accuracy and value of information. The second task is to gain a sense of the current trend in air quality forecast value by assessing the next-generation of operational numerical air quality models. The third task is to develop the forecast tool using the information from the first two tasks as guidelines. The tool must improve upon the value of information already provided by the current and next-generation forecast systems. Each of these tasks will be described following a brief background discussion that further motivates this research.

1.1 Air Quality

Air quality is defined as the condition of the air as it relates to the requirements and preferences of humans and wildlife including health and well being (Johnson et al., 1997). Air pollution, therefore, would cause a significant degradation in air quality which impacts human health and well being. The Clean Air Act of 1970 tasked the U.S. Environmental Protection Agency (EPA) with the responsibility of protecting the nation's air and water. As a result, air pollution levels have dropped, drastically improving air quality. Unfortunately, 42% of the U.S. population still suffers from harmful pollution levels (Nolen et al., 2013).

The EPA has designated six criteria pollutants to be monitored for their effects on human health. Of these criteria pollutants, ozone and particulates regularly exceed healthy limits as outlined in the National Ambient Air Quality Standards (NAAQS). The local concentrations of these pollutants are reported to the general public using an Air Quality Index (AQI). The AQI is a method of normalizing pollutant concentrations into a single number that conveys the level of threat to human health (Mintz, 2009). The AQI is a linear interpolation between pollutant breakpoints shown in Table 1.1 and is often color-coded, enabling quick interpretation of health

Table 1.1: The Air Quality Index (AQI) with associated health risk and color code. Breakpoints in parentheses are for dates prior to March 2008 when the ozone standards changed. Adapted from Mintz (2009).

AQI Range	**Ozone Break Points ( ) = pre-2008**	**Level of Health Concern**	**Color Code**
0 - 50	0 - 59 (0 - 64) ppbv	Good	Green
51 - 100	60 - 75 (65 - 84) ppbv	Moderate	Yellow
101 - 150	76 - 95 (85 - 104) ppbv	Unhealthy for Sensitive Groups	Orange
151 - 200	96 - 115 (105 - 124) ppbv	Unhealthy	Red
201 - 300	116 - 374 (125 - 374) ppbv	Very Unhealthy	Purple
301 - 500	375 - 600 (375 - 600) ppbv	Hazardous	Maroon

risk. Pollutant concentrations producing AQI in the "Unhealthy for Sensitive Groups" or code-orange category are deemed in exceedance of the NAAQS and are typically associated with poor air quality.

1.2 Ozone

Ozone is an abundant oxidant in the atmosphere with roughly 90% residing in the Stratosphere, protecting the surface by absorbing ultra-violet radiation (Seinfeld and Pandis, 2006). Near-surface ozone, however, is a harmful pollutant (Lippmann, 1989; Wright et al., 1990; Berry et al., 1991) and thus will be the focus of this research. This near-surface ozone is produced primarily though reactions with nitrogen oxides (NO_x) and volatile organic compounds ($VOCs$). Below is the photostationary steady state of near-surface ozone (henceforth referred to simply as ozone):

$$\mathrm{NO_2} + h\nu \rightarrow \mathrm{NO} + \mathrm{O} \qquad \text{R1}$$

$$\mathrm{O} + \mathrm{O_2} + \mathrm{M} \rightarrow \mathrm{O_3} + \mathrm{M} \qquad \text{R2}$$

$$\mathrm{O_3} + \mathrm{NO} \rightarrow \mathrm{NO_2} + \mathrm{O_2} \qquad \text{R3}$$

NO_2 is photolyzed in the presence of sunlight (wavelengths < 424 nm), yielding a single oxygen atom. This oxygen atom reacts with diatomic oxygen to produce ozone. Ozone then reacts with the NO from reaction R1 regenerating NO_2 and diatomic oxygen. The M in reaction R2 is simply a non-reactive molecule in the presence of the reaction used to absorb energy (such as N_2). The concentration of ozone in this steady state is thus dependent upon the ratio of the concentrations of NO_2 to NO, but this cycle produces no new ozone.

Reactions of $VOCs$ with the hydroxyl radical (OH) interrupt the photostationary steady state, leading to a net production of ozone. For example, methane (CH_4) reacts with OH in the

presence of diatomic oxygen, yielding a methyl peroxy radical (CH_3O_2) and water:

$$\mathrm{CH_4 + OH \xrightarrow{O_2} CH_3O_2 + H_2O} \qquad \text{R4}$$

In troposhperic conditions, the methyl peroxy radical reacts with NO in the presence of diatomic oxygen to produce formaldehyde ($HCHO$), the hydroperoxyl radical (HO_2), and NO_2:

$$\mathrm{CH_3O_2 + NO \xrightarrow{O_2} HCHO + HO_2 + NO_2} \qquad \text{R5}$$

The NO_2 produced in this reaction is introduced into the photostationary steady state reaction R1 resulting in the net production of ozone. The HO_2 and $HCHO$ can continue to react and form additional ozone.

It is clear from the photostationary steady state that sunlight is needed to produce ozone; thus, the late-spring through early-fall months are considered the ozone season when actinic flux peaks. In the U.S., the ozone season is generally April - October, though some higher latitude locations experience a slightly shorter ozone season.

1.3 Region of Interest

The mid-Atlantic region of the U.S. is the primary region of interest, which includes Washington D.C., Northern Virginia, Baltimore, and surroundings. This region was chosen for a few reasons. This region is hit particularly hard with poor air quality during the ozone season, ranking number 9 on the list of most ozone-polluted regions in the United States (Nolen et al., 2013). While the levels of air pollution in the mid-Atlantic region are comparable to those in southern California or eastern Texas, the mid-Atlantic is often overlooked in the literature. The ozone in this region is well monitored and the state agencies, specifically the Maryland Department of the Environment and the Virginia Department of the Envirionment, have expressed great interest in this research.

1.4 Value of Information

Most forecast systems are developed to aid in making decisions. Information produced from a forecast system would influence the future actions of a decision-maker, such as a forecast for rain may convince the user to carry an umbrella. Low-quality information may lead to decisions yielding sub-optimal results, thus it is important to assess not only the skill of a forecast system, but also the quality of information the forecast yields in a decision-making scenario. This assessment is often referred to as assessing the value of information.

A simple, yet powerful, method of assessing the value of information is through a cost-loss ratio model (Katz and Murphy, 1997; Richardson, 2000; Wilks, 2011; Garner and Thompson, 2012). Consider the following example. The forecast for tomorrow is that there is a chance of rain. Your decision is whether or not to carry an umbrella. The contingency table associated

Table 1.2: Contingency table indicating the cost of an action (C) that protects against a loss (L) depending on the observed state of the atmosphere. There is neither cost nor loss due to inaction on non-event days (i.e. it does not rain or there is good air quality).

	Action	**No Action**
Non-Event	C	—
Event	C	L

with this decision is shown in Table 1.2. Carrying an umbrella would come at some minor cost C, which doesn't change if it does or does not rain. You break even if you don't carry an umbrella and it doesn't rain; however, you incur a loss L (i.e. you get soaked) if you don't carry an umbrella and it rains. This concept can be easily applied to forecasting air quality. Giving free bus rides or reducing power plant output, depending on an air quality forecast, act as the "umbrella" that protects against effects of poor air quality, such as increased emergency room visits and loss of tourism. It is the value of air quality forecasts in these and similar decision scenarios that is the focus of this work.

Additional introductory and background information is available in the individual chapters to follow. Chapter 2 covers the assessment of the current state of value of three different air quality forecast systems. Chapter 3 is an analysis of numerical air quality models in forecasting high ozone events in the mid-Atlantic during a NASA field campaign. Chapter 4 describes the development and evaluation of the ensemble statistical post-processor. Chapter 5 summarizes the previous chapters under the scope of the dissertation as a whole.

Chapter 2

The Value of Air Quality Forecasting in the Mid-Atlantic

2.1 Introduction

Air quality has been of growing interest in both the public eye and the scientific community. Air pollution has been extensively studied since the 1960s where, in California, urban growth and unique terrain are conducive to air pollution problems. Identifying these problems spawned work on predictive methods (McCollister and Wilson, 1975; Aron and Aron, 1978) in an attempt to forecast high pollution episodes, protect public health and environment aesthetics, and provide information regarding air pollution policy.

Health studies such as Krupnick and Ostro (1990) and Berry et al. (1991) indicate that prolonged exposure to air pollution can cause ocular irritation, reduced lung function, and general degradation of the respiratory system. It is only natural that the health-conscious public grows concerned as they become aware of the risks associated with polluted air. This, in turn, puts pressure on air shed managers to help mitigate emission of and exposure to air pollutants and their precursors. Strategies and programs aimed at reducing emissions and exposure to pollutants can be costly endeavors; therefore, air shed managers need forecasts to help them make accurate and timely decisions. These forecasts are often chosen based on their performance as measured by statistical quantities such as correlation, bias, and error. This study critically evaluates the sources of these forecasts by going beyond the standard evaluation methods and incorporating the decision-making aspects associated with air quality forecasting.

2.1.1 Air Quality Forecast Systems

This paper focuses on assessing the value of forecasts produced by an operational numerical model and forecasts produced by human forecasters. The value of persistence forecasts are assessed for relative comparison. The model, human forecasters, and persistence forecasting are collectively

referred to as forecast systems.

2.1.1.1 The National Air Quality Forecast Capability

The National Oceanic and Atmospheric Administration (NOAA) and the Environmental Protection Agency (EPA) developed the National Air Quality Forecast Capability (NAQFC, also known as the National Air Quality Forecast System [NAQFS] in previous literature) in partial fulfillment of the Energy Policy Act of 2002. The NAQFC couples the Weather Research and Forecasting Non-Hydrostatic Mesoscale Model (WRF-NMM) (Janjic, 2003) with the Community Multiscale Air Quality (CMAQ) model (Byun and Schere, 2006) to produce 48-hour forecasts of surface 1-hour average and 8-hour average ozone mixing ratios across the contiguous U.S., Alaska, and Hawai'i. The NAQFC also produces forecasts for smoke and will soon be adapted to produce forecasts for particulates; however, these will not be discussed here. Current NAQFC data are available through the National Weather Service (NWS) National Digital Guidance Database (http://www.weather.gov/aq) and archived NAQFC data are available through NOAA's National Operational Model Archive and Distribution System (NOMADS).

The NAQFC has been subjected to rigorous verification (Ryan et al., 2004). Eder et al. (2006) performed the standard model evaluation statistics, such as bias and error analyses, in part to improve the NAQFC performance and prepare the NAQFC for operational use in 2005. Eder et al. (2009) evaluated the model using both standard and categorical statistics, such as false alarm rates, hit rates, and critical success indices. Eder et al. (2010) use areal forecast statistics in the latest NAQFC verification rather than point-to-point comparisons as done in the previous evaluations. These categorical statistics assess the model performance in a more typical air quality forecasting scenario where a forecast is issued for an area rather than a specific site. Their results indicate that the NAQFC forecast skill is comparable to that of expert human air quality forecasters and that the NAQFC will become more useful under stricter air quality standards.

2.1.1.2 Human Forecasters and Persistence Forecasting

Air quality forecasters typically develop their own tools and algorithms to forecast pollution episodes in their respective areas of interest (Aron and Aron, 1978; Lin, 1982; Robeson and Steyn, 1990; Ryan, 1995; Hubbard and Cobourn, 1998; Davis and Speckman, 1999). Human air quality forecasters, however, offer real-world experience in blending standard information and novel interpretations of non-traditional data, such as human behavior and current events, into forecast algorithms.

A human air quality forecast for the following day in the mid-Atlantic region is submitted to AIR-Now Tech by 1500 local time. The time to prepare a forecast can be limited due to the forecaster's responsibilities outside of producing an air quality forecast. The ozone forecast is for the maximum 8-hour average surface ozone mixing ratio for the day. A forecaster uses a wide array of data and information to create a forecast, and each forecaster has many preferred

sources of information. Although there is no regular procedure to use, many forecasters often use similar sources. A human air quality forecast would typically include an analysis of the most recent meteorological-model simulations, air quality observations from sites within the forecast region and neighboring areas, and oftentimes a personally developed statistical model or empirical forecast rule (W. F. Ryan 2009, personal communication). The human forecasters do have access to the NAQFC when producing a forecast; however, it is one source of information, among many, that the human forecaster considers.

Persistence forecasting is a simple and straight-forward forecast method. The idea is to use the current conditions as a forecast for the next day. Persistence forecasting is often regarded as a primitive forecast method, but it can be useful under specific (i.e. persistent, hence the name) circumstances. Poor air quality events are typically episodic, which is why persistence can be considered a good benchmark forecast which is then adjusted according to information at hand.

2.1.2 Air Quality and Decision Making

2.1.2.1 Ozone and the Air Quality Index

Ozone is an abundant oxidant in the atmosphere. Near-surface ozone, however, has been extensively studied and found to be a harmful pollutant (Lippmann, 1989; Wright et al., 1990; Berry et al., 1991; Chen et al., 2007). This ozone is produced by the photolysis of nitrogen oxides (NO_x) and oxidation of volatile organic compounds (VOCs) (Seinfeld and Pandis, 2006). NO_x is primarily produced by combustion processes such as natural fires, vehicles, and power-plants. VOCs are produced by plant life, burning of carbon-based fuels, and in thousands of industrial processes such as plastics manufacturing and smelters. The EPA designated ozone as a criteria pollutant because of its heavy dependence on anthropogenically produced precursors and malicious effects on human health. As a criteria pollutant, ozone is subject to restrictions outlined in the National Ambient Air Quality Standards (NAAQS [http://www.epa.gov/air/criteria.html]).

Currently, six criteria pollutants are monitored by state agencies and reported to the EPA as part of the Clean Air Act. The concentration of these criteria pollutants is made available to the general public via an Air Quality Index (AQI). As discussed in Mintz (2009), the AQI is a method of normalizing pollutant concentrations into a single number to convey the level of threat to public health. The AQI is calculated using a standard linear interpolation from the pollutant concentration to the index using the break points in Table 2.1. For convenience, the AQI is color coded to enable quick interpretation of the health risk. Poor air quality is generally associated with an orange AQI (AQI = 101) or higher. Of the criteria pollutants, particulates and ozone routinely reach unhealthy levels in many urban areas throughout the US. This study will focus on ozone in the Mid-Atlantic region where elevated ozone is often observed, and a variety of protective measures are employed by local and regional governments.

Table 2.1: The Air Quality Index (AQI) with associated health risk and color code. Breakpoints in parentheses are for dates prior to March 2008 when the ozone standards changed. Adapted from Mintz (2009).

AQI Range	**Ozone Break Points ( ) = pre-2008**	**Level of Health Concern**	**Color Code**
0 - 50	0 - 59 (0 - 64) ppbv	Good	Green
51 - 100	60 - 75 (65 - 84) ppbv	Moderate	Yellow
101 - 150	76 - 95 (85 - 104) ppbv	Unhealthy for Sensitive Groups	Orange
151 - 200	96 - 115 (105 - 124) ppbv	Unhealthy	Red
201 - 300	116 - 374 (125 - 374) ppbv	Very Unhealthy	Purple
301 - 500	375 - 600 (375 - 600) ppbv	Hazardous	Maroon

2.1.2.2 Costs associated with air quality

Air shed managers need to consider a variety of direct and indirect costs associated with air quality. When ozone is forecasted to be above the NAAQS threshold (AQI $>$ 100) on a given day, an "Alert Day" may be invoked with special programs. The implementation of such programs is designed to reduce the emissions of the ozone precursors and avoid losses due to high levels of ozone such as health costs and EPA non-attainment (Anderson, 2001). Accurate ozone forecasts are needed to make proper decisions regarding program implementation. Inaccurate forecasts could lead to unneeded action or missed opportunities to mitigate ozone precursor emissions, both of which would result in preventable monetary loss.

2.2 Data

Ozone observations were assembled from the Air Quality System (AQS), a database hosted by the EPA Technology Transfer Network with contributions from local, state, and federal agencies. The data specifically used in this study were provided to the AQS by the Virginia Department of Environmental Quality (VADEQ), the Maryland Department of the Environment (MDE), and the District Department of the Environment (DDOE). Average hourly surface ozone mixing ratios were collected for the 2005 - 2009 ozone seasons (April - October) at 40 locations throughout Maryland, Washington DC, and Virginia. The 1-hr averages were converted to hourly 8-hr forward running averages which were then used to calculate the AQI. The breakpoints for calculating ozone AQI were reduced in 2008, therefore AQI were calculated using the old breakpoints for the 2005 - 2007 ozone seasons and the new breakpoints in the 2008 and 2009 ozone seasons. The maximum AQI for a given day among all the sites within a region would represent the observed AQI for that region on that day similar to the methods used in the second evaluation approach in Eder et al. (2010). Refer to Fig. 2.1 for a map color-coded by region. NAQFC forecast 8-hr average surface ozone mixing ratios were obtained from NOMADS for the same time period and sites as the observations and were converted to AQI. The maximum AQI out of all the sites

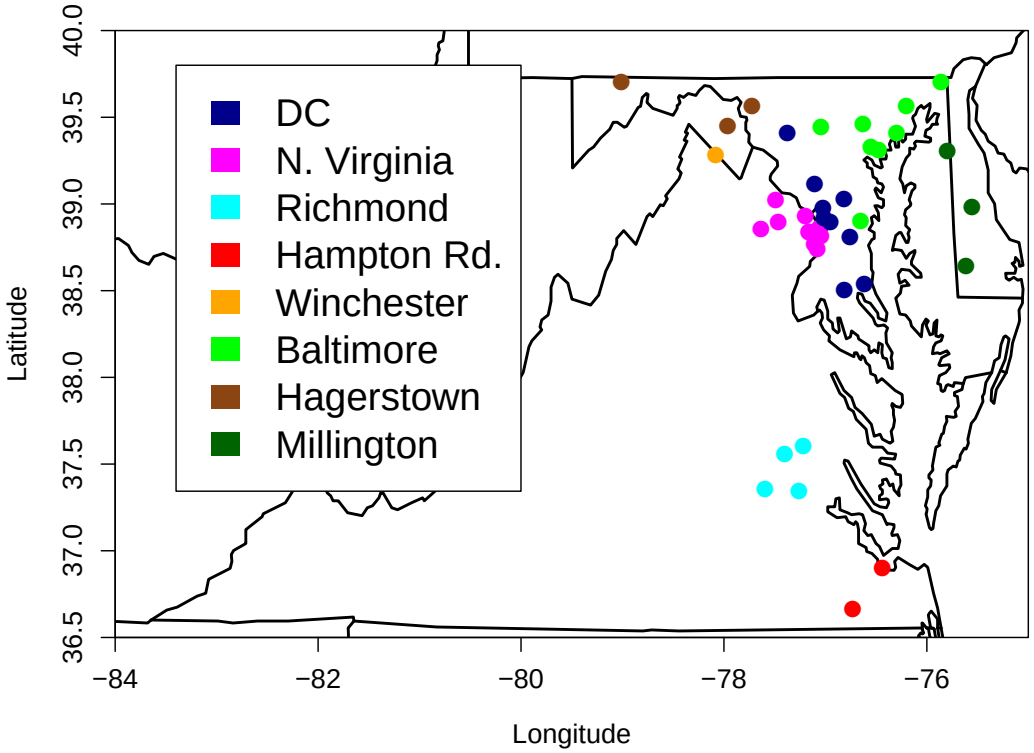

Figure 2.1: A map of the sites within each forecast region. The dots represent individual sites and the colors represent forecast regions. Note that some sites do overlap.

within a region is used for the analysis. Human air quality forecasts of daily maximum AQI were also obtained directly from the VADEQ and MDE for the same time period. These forecasts were provided for the regions, not individual sites, so no further conversions or calculations were needed. Persistence forecasts were developed simply by using the maximum AQI for the current day as the forecast for the next day. Days missing either an observation or a forecast for any one of the forecast systems were excluded from this study.

2.3 Methods

Standard verification statistics were computed to assess the skill of each forecast system. We used those verification statistics set forth by Eder et al. (2010), including the mean bias (MB), normalized mean bias (NMB), root-mean-square error (RMSE), normalized mean error (NME), and correlation coefficient (R) of each forecast system as a whole, and categorical statistics such as critical hit rate (cH), exceedance hit rate (eH), exceedance false alarm rate ($eFAR$), and exceedance critical success index ($eCSI$) for each forecast system within each region. These categorical statistics are explained in greater detail in Eder et al. (2010).

Table 2.2: Contingency table indicating the cost (C) to protect against a loss (L) depending on the state of the air quality. There is neither cost nor loss when no protective measure is taken on a good air quality day.

	Protect	**Do Not Protect**
Good AQ	C	-
Poor AQ	C	L

A static cost-loss ratio model was used to assess the value of each forecast system. This model was first discussed in a meteorological context in Thompson (1952) and Thompson and Brier (1955). Kernan (1975) is an early example of applying the model to an air pollution decision-making situation. The cost-loss ratio model is discussed in detail in the Thompson papers, the Kernan paper, and more recently in Katz and Murphy (1997), Richardson (2000), Thornes (2001), and Berger (2006). Recent examples of the application of the cost-loss ratio model can be found in Rotach et al. (2009) and Millner (2009).

The method used to calculate value in this study is detailed in Richardson (2000) and is summarized here. Start with a 2x2 contingency table that compares the cost (C) to protect or insure against a loss (L) if a poor air quality event were to occur (Table 2.2). C and L are often expressed as some form of currency. Examples of C include free bus rides, reduced production at power-plants, and other measures aimed at reducing pollutant and pollutant precursor concentrations. Examples of L include loss of tourism, environmental degradation, and an increased number of patients in hospital emergency rooms. Insuring against a loss will cost C, whether or not the poor air quality event occurs; however, one would lose L if the poor air quality event occurs and no protective measures were implemented. One assumption in this model is that $C < L$, which makes sense, for there would be no decision to make otherwise.

Now consider N number of cases, each case with an air quality forecast and observation. N_f is the number of times the event was forecasted, but not observed, and N_o is the number of times the event was observed, but not forecasted. N_{fo} is the number of correctly forecasted poor air quality events. The climatological frequency of these events is defined as $s = \frac{N_{fo}+N_o}{N}$. Expected losses can be formulated using Table 2.2 and the definitions above.

The expected loss using no forecast system is

$$E_c = min(C, sL) \tag{2.1}$$

where the air shed manager chooses the minimum expense by deciding whether to always protect at a cost of C or never protect and lose sL. The expected loss using a forecast system is

$$E_f = \frac{1}{N}(N_{fo}C + N_fC + N_oL) \tag{2.2}$$

where the air shed manager would protect whenever a poor air quality event is forecasted and

not protect otherwise. The expected loss of a perfect forecast system is

$$E_p = sC \tag{2.3}$$

where the only expense incurred is the cost to protect when a poor air quality event occurs. Using Eq. (2.1) - (2.3), the relative value of the forecast system can be defined as

$$V = \frac{E_c - E_f}{E_c - E_p}. \tag{2.4}$$

Although estimates of C may be available, it is very difficult to estimate L over many environmental and social aspects, making it difficult to quantify value for a specific protective measure. By dividing both the numerator and denominator of Eq. (2.4) by L, the relative value can be expressed in terms of a cost-loss ratio ($\frac{C}{L}$) as

$$V = \frac{min(\frac{C}{L}, s) - \frac{1}{N}[N_{fo}\frac{C}{L} + N_f\frac{C}{L} + N_o]}{min(\frac{C}{L}, s) - s\frac{C}{L}}. \tag{2.5}$$

This expression allows value to be quantified for a broad range of protective measures without having to know C and L exactly. Value is undefined at cost-loss ratios equal to zero and one ($\frac{C}{L} = 0$ or $1, V = undefined$). This approach is consistent with a decision situation. There would be no decision to make when there is no cost to protect ($C = 0$, always protect) or when the cost to protect is equal to the loss incurred ($C = L$, never protect); therefore, the value of the forecast can not be determined. The value of a perfect forecast system (i.e. $E_f = E_p$) would be unity at cost-loss ratios between zero and one ($0 < \frac{C}{L} < 1, V = 1$). For non-perfect forecasts, the value will range from zero to one. For this study, a poor air quality event is defined as a case when the observed AQI is greater than 100. This is to remain consistent with the NAAQS threshold for an "Air Quality Alert" day.

2.4 Forecast System Verification

2.4.1 Overall

The scatter plots in Fig. 2.2 depict the discrete statistics and overall skill of the human forecaster (top-left), the NAQFC (top-right), and persistence (bottom). The human forecaster out-performs both the NAQFC and persistence according to the correlation coefficient (0.73 compared to 0.65 and 0.58 respectively), RMSE (18.3 compared to 25.2 and 23.4), and NME (23.9% compared to 33.3% and 30.1%). The NAQFC is heavily positively biased (MB of 10.97 and NMB of 20.57%) whereas both the human forecasters and persistence forecasting are only slightly positively biased. Although the human forecasters have a slight positive bias, a preponderance of the observation-forecast pairs fall below the 1:1 line indicating conservative forecasting.

Confidence intervals were calculated on the R, RMSE, and MB discrete statistics using a

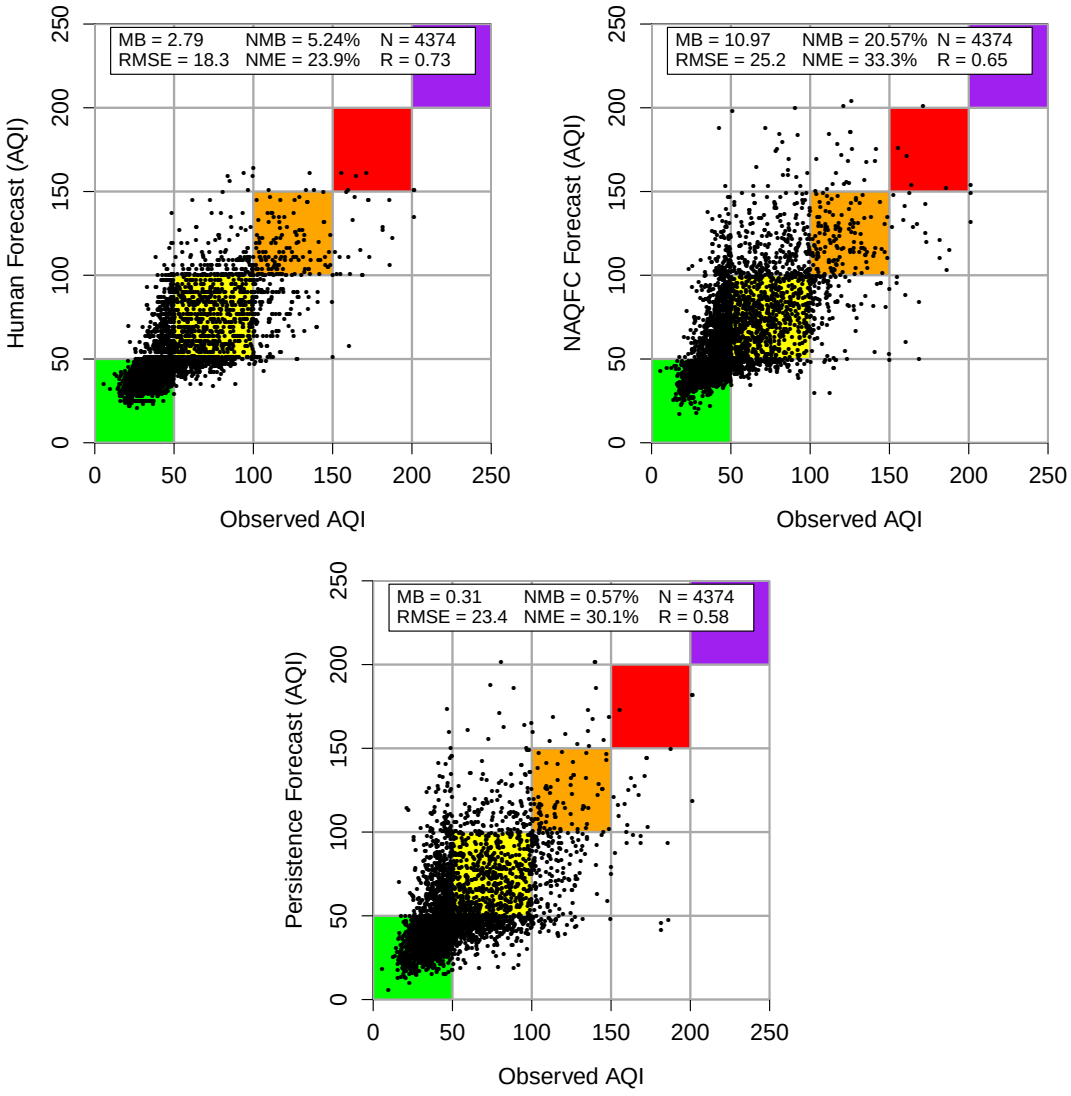

Figure 2.2: Scatter plots with discrete statistics for the human forecaster (top-left), NAQFC (top-right), and persistence forecasting (bottom) over all the forecast regions. Each dot represents an observation-forecast pair.

bootstrap algorithm. The observation-forecast pairs were repeatedly sub-sampled with replacement, creating 10,000 bootstrap samples from which the 95% confidence intervals were derived. The distributions and confidence intervals of these bootstrap samples are shown in Fig. 2.3. The horizontal lines under the distributions indicate the 95% confidence interval. None of the confidence intervals overlap, implying that the differences between the forecast systems' R, RMSE, and MB are statistically significant at 95%.

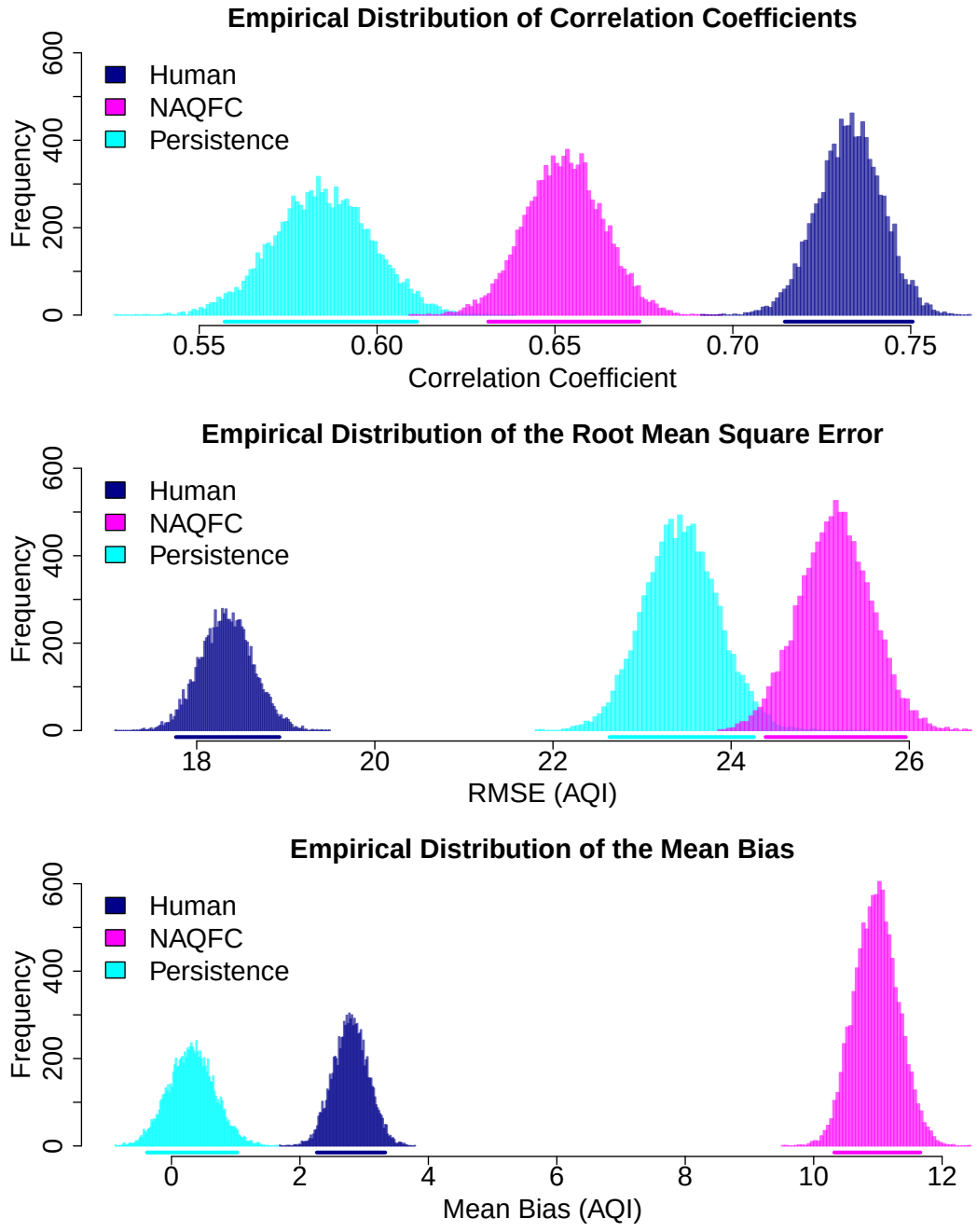

Figure 2.3: Empirical distributions with 95% confidence intervals of the correlation coefficients (top), root-mean-square error (middle), and mean bias (bottom) of each forecast system. Each distribution was developed using 10,000 bootstrap sub-samples of observation-forecast pairs.

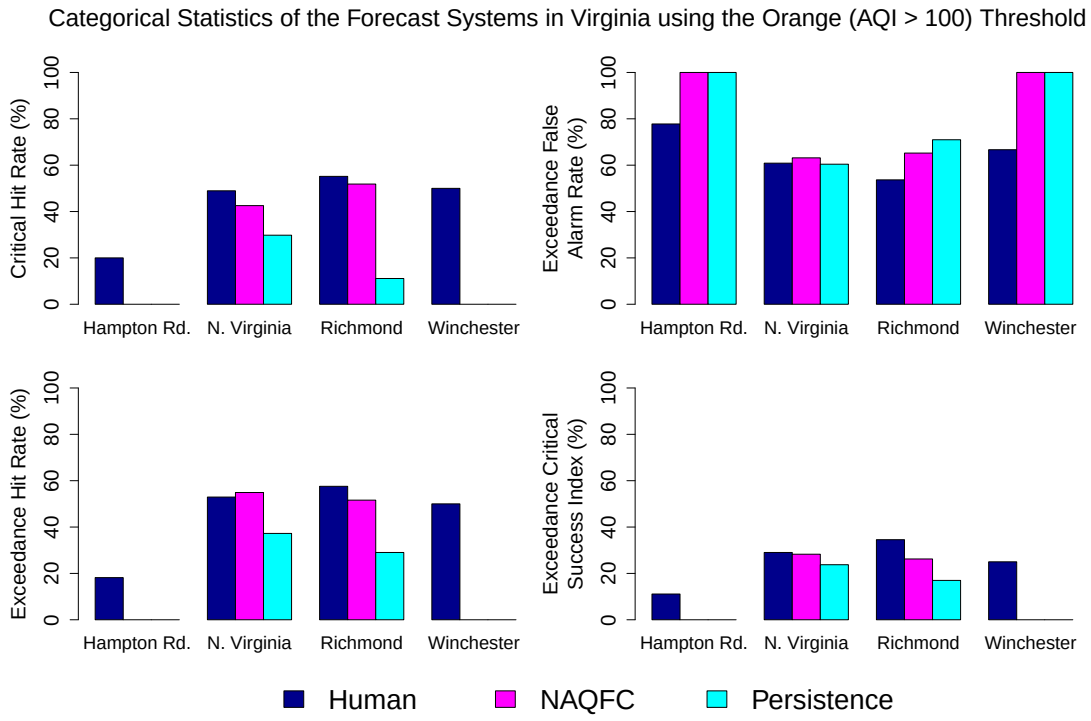

Figure 2.4: Categorical statistics for all of the Virginia regions. Statistics include, from top-left to bottom-right, the critical hit rate, exceedance false alarm rate, exceedance hit rate, and exceedance critical success index.

2.4.2 Regional and Threshold Specific

Categorical statistics were calculated for each forecast system within each region at the orange AQI threshold. Fig. 2.4 describes the categorical statistics for the Virginia regions. The four panels show, from top-left to bottom-right, the cH, $eFAR$, eH, and $eCSI$ for each region within the state. Hampton Roads and Winchester recorded fewer exceedance days than the other six regions (see Table 2.3), explaining the low cH, eH, $eCSI$, and the high $eFAR$ in Fig. 2.4. The human forecaster performs better than both the NAQFC and persistence in all regions and statistics except one. The NAQFC performed better according to the eH in Northern Virginia; however, the strong positive bias in the NAQFC gives the model an advantage in this statistic. The human forecaster $eCSI$ is only slightly higher than both the NAQFC and persistence $eCSI$ in Northern Virginia and approximately 10% - 20% higher in Richmond. The high $eFAR$ for both the NAQFC and persistence dramatically reduced their respective $eCSI$ in these regions. The $eCSI$ for the NAQFC and persistence in both Hampton Roads and Winchester is zero because the eH for both these forecast systems is zero in these regions.

The categorical statistics for the Maryland regions are depicted in Fig. 2.5 with a similar setup to Fig. 2.4. The NAQFC and persistence forecasting performs the same (within 2%) if not better than the human forecaster in both the hit rate statistics in all of the Maryland

Table 2.3: Frequency chart used in categorical statistics calculations using the orange AQI threshold (AQI $>$ 100). N is the total number of observation-forecast pairs. N_o is the number of observations not forecasted above the threshold. N_f is the number of forecasts not observed over the threshold, and N_{fo} is the number of observations forecasted above the threshold.

		Human			**NAQFC**			**Persistence**		
	N	N_o	N_f	N_{fo}	N_o	N_f	N_{fo}	N_o	N_f	N_{fo}
N. Virginia	460	24	39	27	23	48	28	32	29	19
Richmond	460	13	20	18	15	30	16	22	22	9
Hampton Rd.	461	8	7	2	10	11	0	10	10	0
Winchester	465	1	2	1	2	7	0	2	3	0
DC	628	34	37	38	17	114	55	43	43	29
Baltimore	686	39	33	55	35	76	59	55	58	39
Hagerstown	648	10	4	2	10	10	2	9	8	3
Millington	566	11	13	10	8	78	13	13	14	8

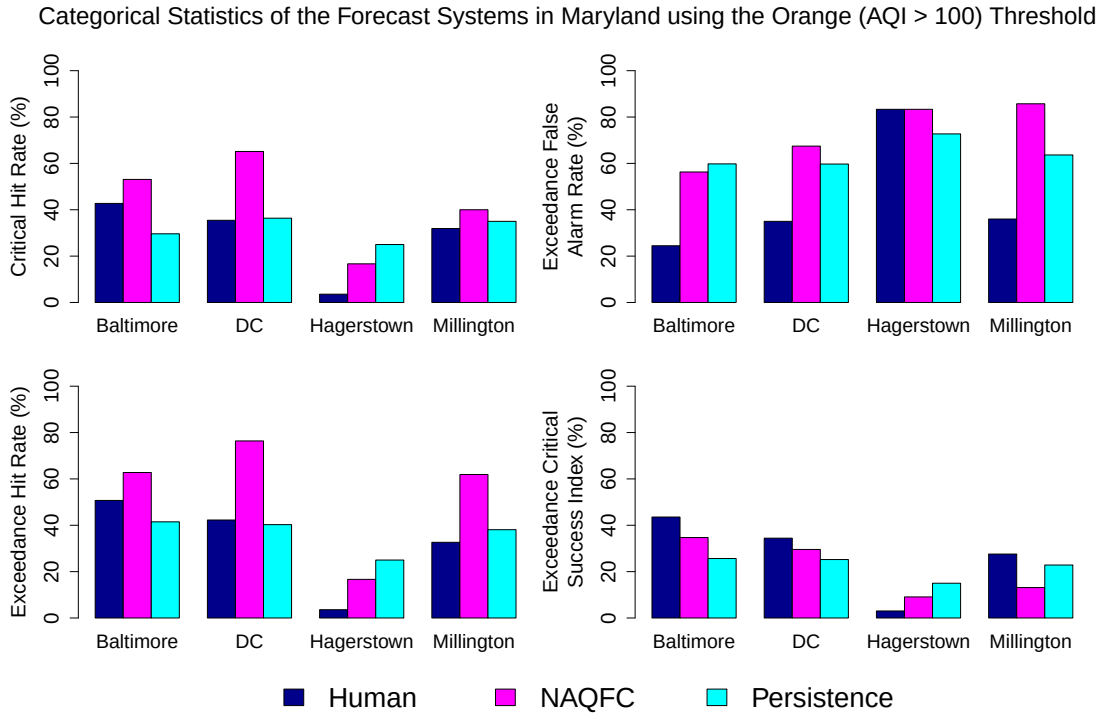

Figure 2.5: Categorical statistics for all of the Maryland regions. Statistics include, from top-left to bottom-right, the critical hit rate, exceedance false alarm rate, exceedance hit rate, and exceedance critical success index.

regions except Baltimore. Just as in the Virginia regions, the high $eFAR$ for both the NAQFC and persistence forecasting drastically reduces their respective $eCSI$. The $eCSI$ indicates that the human forecaster outperforms the NAQFC and persistence forecasting in Baltimore, DC, and Millington by 5% - 15%; however, persistence forecasting performs the best in Hagerstown. This could be explained by the infrequent observations of ozone above the orange AQI threshold (Table 2.3) in Hagerstown and that one of the verification sites within the region is at an elevation of 764 m, making accurate forecasts difficult.

2.4.3 Value of Forecast Systems

The value of each forecast system was calculated with Eq. (2.5) at the orange AQI threshold and are discussed in this section. The value of each forecast system was also calculated for the red and yellow AQI thresholds, but are not included in this study. There were too few days with observed ozone in the red AQI range to make a robust statistical analysis. Conclusions drawn from the yellow threshold analysis were consistent with the orange threshold analysis. Including these results would be redundant and have little meaning since alert days are rarely, if ever, invoked on a code yellow day.

Fig. 2.6 shows the value curves for all the forecast regions in both Virginia and Maryland using the human forecaster (top-left), the NAQFC (top-right), and persistence (bottom-left). The calculated value lies on the ordinate and the ratio of the costs to losses lies along the abscissa of each of the value plots in Fig. 2.6. The value can be interpreted as the percent saved from the difference between no forecast system and a perfect forecast system. Assuming L does not change, the cost-loss ratio may be interpreted as a given protective measure to insure against the loss. For example, the Baltimore value curve for the human forecaster peaks at 0.529 with a cost-loss ratio of 0.137 (Table 2.4). This means that the air shed manager would save 52.9% of the difference between the expense of no forecast system and a perfect forecast system when deciding to issue a protective measure costing 13.7% of the losses. If the losses total \$100,000 in this situation, one would expect to save \$6,254 per forecasted event ($13.7\% \times [\$100,000 - \$13,700] \times 52.9\% = \$6,254$).

The value curves of the human forecasts cover a broader range of cost-loss ratios than either the NAQFC or persistence forecast value curves. Although the minimum cost-loss ratio at which each forecast system has positive value is similar among all the regions (between 0.01 and 0.1), the maximum cost-loss ratio over which the human forecast has positive value is 10% - 30% higher than the NAQFC and persistence forecasts for their respective regions. This indicates that the human forecaster produces forecasts that can save air shed managers money over a wider range of protective measures than the other forecast systems. Among the forecast systems, the NAQFC forecasts produce the highest peak-value in DC ($V = 0.558$) and Millington ($V = 0.475$) while persistence forecasts produce the highest peak-value in Hagerstown ($V = 0.237$). The human forecasts produce the highest peak-value in the remaining regions. Neither the NAQFC nor persistence forecasts produce value by our definition at Hampton Roads and Winchester because neither forecast system correctly forecasted a poor air quality event during the analysis period

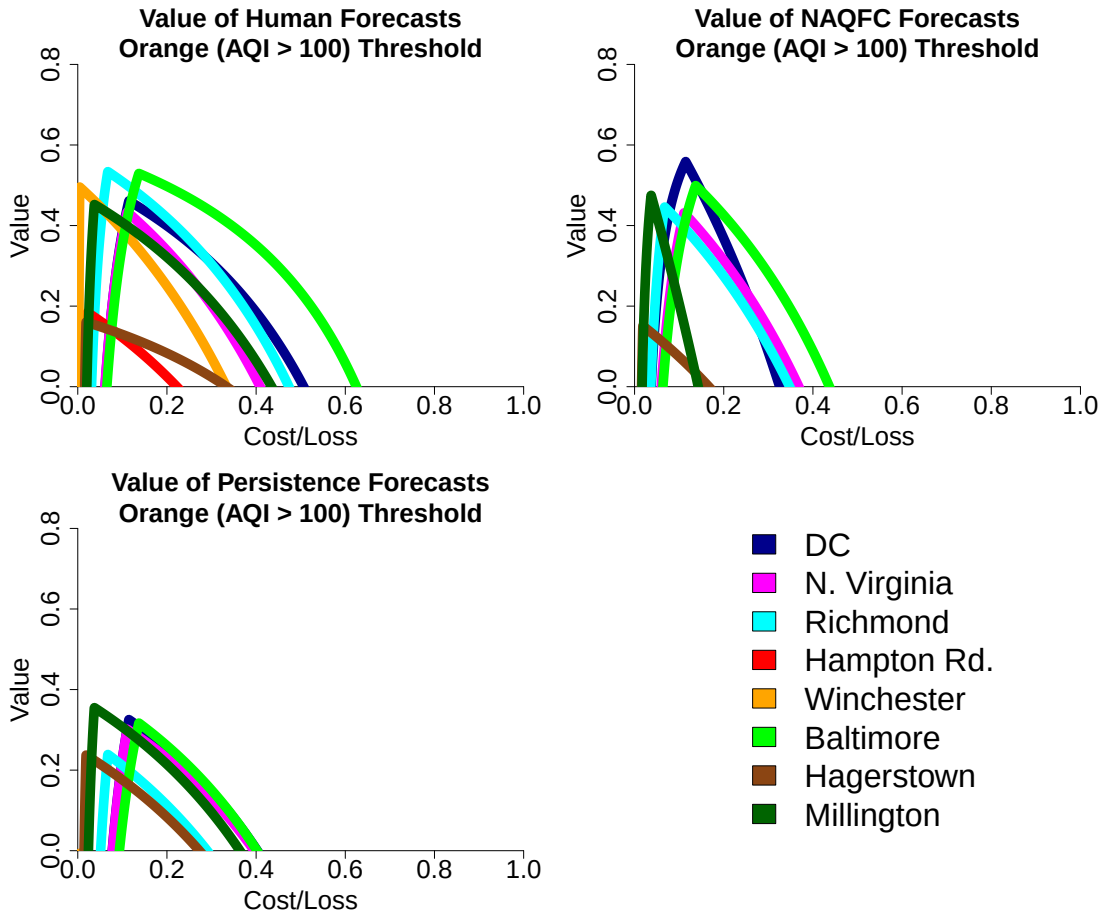

Figure 2.6: Value curves for the human forecaster (top-left), the NAQFC (top-right), and persistence forecasting (bottom-left). Value curves were calculated using an orange AQI threshold (AQI $>$ 100).

($eH = 0$).

With access to the three forecast systems discussed in this study, the air shed manager would be able to choose the forecast system that produces the highest value depending on the protective measure. The plots in Fig. 2.7 show the maximum value of the three forecast systems at each region as a function of the cost-loss ratio. These curves are a combination of the human, NAQFC, and persistence value curves and are color-coded according to the forecast system that produces the highest value for that specific cost-loss ratio. Of the eight forecast regions, five benefit from a combination of forecast systems. In the DC region, the NAQFC produces the highest value up to a cost-loss ratio of 0.2, after which the human forecaster produces the highest value. In the Hagerstown region, persistence forecasts produce the highest value up to a cost-loss ratio of 0.3, after which the human forecast produces the highest value. The NAQFC technically produces the highest value at low cost-loss ratios in the Northern Virginia, Baltimore, and Millington regions; however, the combined curves are not much different than the human forecast value curves.

Table 2.4: Table showing the peak value, the cost-loss ratio of the peak value, and the range of cost-loss ratios over which each forecast system holds value.

	Human		
	Max Value	C/L of Max Value	C/L Range
N. Virginia	0.434	0.111	0.061 - 0.409
Richmond	0.534	0.068	0.031 - 0.473
Hampton Rd.	0.184	0.022	0.018 - 0.222
Winchester	0.495	0.005	0.003 - 0.333
DC	0.461	0.115	0.062 - 0.506
Baltimore	0.529	0.137	0.066 - 0.624
Hagerstown	0.160	0.019	0.016 - 0.333
Millington	0.451	0.038	0.021 - 0.434
	NAQFC		
	Max Value	C/L of Max Value	C/L Range
N. Virginia	0.432	0.111	0.060 - 0.368
Richmond	0.446	0.068	0.037 - 0.347
Hampton Rd.	—	—	—
Winchester	—	—	—
DC	0.558	0.115	0.038 - 0.325
Baltimore	0.499	0.137	0.064 - 0.437
Hagerstown	0.151	0.019	0.016 - 0.166
Millington	0.475	0.038	0.017 - 0.142
	Persistence		
	Max Value	C/L of Max Value	C/L Range
N. Virginia	0.302	0.111	0.078 - 0.395
Richmond	0.239	0.068	0.052 - 0.290
Hampton Rd.	—	—	—
Winchester	—	—	—
DC	0.325	0.115	0.078 - 0.402
Baltimore	0.317	0.137	0.094 - 0.402
Hagerstown	0.237	0.019	0.015 - 0.272
Millington	0.355	0.038	0.024 - 0.363

These results indicate that forecast skill does not directly translate to forecast value. The choice of forecast system is highly dependent upon the user's particular need, and using an overall measure of skill, such as $eCSI$, RMSE, R, etc., can often hide this. In some decisions, the less-skillful forecast system may provide the highest value.

2.5 Summary and Conclusions

Standard discrete and categorical statistics were used to evaluate human air quality forecasts, the NAQFC forecasts, and persistence forecasts in eight regions throughout Virginia and Maryland

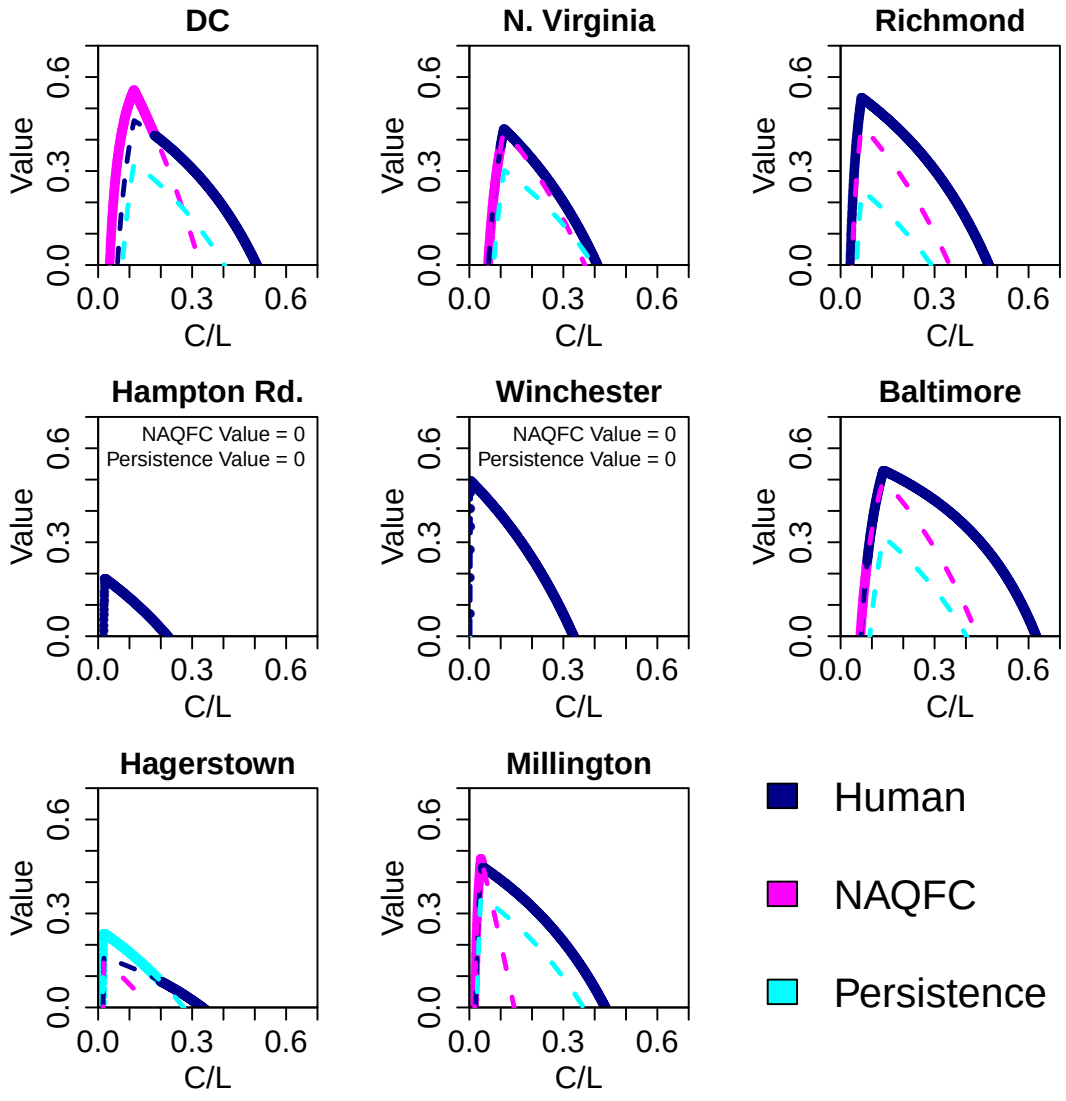

Figure 2.7: Value curves for each forecast system separated by forecast region. The lines are color-coded according to the forecast system. The dashed lines indicate the value of each forecast system. The bold lines indicate the forecast system with the highest value at the given cost-loss ratio.

over five ozone seasons (2005 - 2009). These statistics were supplemented with an assessment of value which incorporates a decision-making aspect into the forecast. The human forecasts performed better than both the NAQFC and persistence forecasts in the discrete statistics. The human forecaster, though slightly positively biased, typically produced conservative forecasts whereas the NAQFC would consistently over-forecast. The categorical statistics indicated that the human forecaster is the most skillful in all the regions except Hagerstown where persistence forecasting is the most skillful. This is likely due to the elevation of the Hagerstown region and

the infrequent occurrences of poor air quality events.

The value of each forecast system varies greatly upon the protective measure being considered. The NAQFC is able to produce higher-value forecasts than human forecasters at relatively low cost-loss ratios in urban and down-wind regions; however, the human forecaster is able to produce high-value forecasts for a broader range of cost-loss ratios in all regions in this study. This implies that the most skillful forecast system may not provide the best value in all situations and thus it is wise for air shed managers to consider multiple forecast systems when deciding on a number of protective measures.

These value metrics can be easily applied to other areas prone to poor ozone events, such as Houston or Los Angeles. These metrics are also being considered to evaluate optimum placement of monitoring sites within forecast regions. A value assessment on individual sites within a region may provide local air shed managers with useful information on their monitoring strategies. A series of case studies performed in a single forecast region, such as Baltimore or DC, that attempt to monetarily quantify losses incurred during a poor air quality event will help tailor these value calculations to more specific decisions and assess the forecast systems' value more accurately.

Chapter 3

Evaluation of NAQFC Model Performance in Forecasting Surface Ozone during the 2011 DISCOVER-AQ Campaign

This chapter is published in the Journal of Atmospheric Chemistry (Garner et al., 2013). It is available with open access through Springer Publishing at `http://dx.doi.org/10.1007/s10874-013-9251-z`.

3.1 Introduction

The NASA Earth Venture Program on **D**eriving **I**nformation on **S**urface conditions from **CO**lumn and **VER**tically resolved observations relevant to **A**ir **Q**uality (DISCOVER-AQ) used a combination of aircraft and ground stations to assess the air quality around Baltimore, Maryland, in July 2011. A team was tasked with providing meteorological and air quality forecasts in support of safe and project-effective flight operations for the two research aircraft. The team provided briefings, which included detailed 24-h forecasts, extended 5-day outlooks, and flight recommendations. With the cost of flight-hours reaching $45,000 per day, it is important to optimize flight decisions using state-of-the-art prognostic tools.

Among the numerous sources of information and tools used to prepare these briefings were two numerical air quality models. One of the numerical models was the National Air Quality Forecast Capability (NAQFC), the current national operational air quality model that provides forecasts of surface ozone and smoke. The other numerical model was an experimental version of the NAQFC (NAQFC-β), which provided forecasts of particulate matter in addition to the current capabilities of the operational model. The NAQFC-β was provided by the NOAA Air

Resources Laboratory to help address a secondary objective of the DISCOVER-AQ campaign, which was to evaluate state-of-the-art air quality models. The combination of these air quality models and the unique flight-decision support needed during DISCOVER-AQ yielded an ideal test-bed for addressing the value of information in real-time decision scenarios.

The questions this work seeks to answer pertain to the secondary objective of DISCOVER-AQ, model evaluation. First, how do each of the numerical air quality models perform during the DISCOVER-AQ campaign? Second, how do the underlying differences between the two models impact the forecasted air quality? Finally, what do these differences mean to the end user in terms of the value of information?

3.2 Background

The NAQFC is an offline system consisting of the Weather Research and Forecasting Non-hydrostatic Meso-scale meteorological Model (WRF-NMM) (Janjic, 2003) coupled with the Community Multi-scale Air Quality Model (CMAQ) (Byun and Schere, 2006) that has been providing forecasts of near-surface ozone since 2005. Forecasts produced by the NAQFC have been rigorously verified with observations (Ryan et al., 2004; Eder et al., 2006, 2009, 2010). The NAQFC is generally biased high when forecasting summertime surface ozone in urban areas of the eastern United States by as much as 5.5 ppbv. Despite the bias, the NAQFC was found to produce valuable forecasts for relatively low to moderate cost decision scenarios, especially in Washington, DC, and Baltimore, MD (Garner and Thompson, 2012).

The NAQFC-β differs from the NAQFC by upgrading CMAQ from a gas-phase-mechanism (version 4.5) to a full gas and aerosol mechanism (version 4.6) while remaining similar in all other aspects (Lee and Ngan, 2011). CMAQ model version 4.6 is based on the Carbon Bond 2005 (CB05) gas-phase chemical mechanism (Yarwood et al., 2005) with modal size-distributed aerosol components (Foley et al., 2010). Both the NAQFC and NAQFC-β are run with the same meteorological fields, emissions inventories, and horizontal grid spacing of 12 km.

Area and mobile emissions are based on the EPA National Emissions Inventory (NEI) for 2005. For Electric Generating Unit (EGU) point sources, Continuous Emission Monitoring 2009 replaces the 2005 NEI where applicable. Updated EGU emissions are further projected into 2011 using emission projection factors from the Department of Energy 2011 Annual Energy Outlook report (`http://www.eia.gov/forecasts/archive/aeo11`). All emissions that are independent from meteorological conditions are processed prior to the model execution using a modified version of the Sparse Matrix Operator Kernel Emission (SMOKE) model (Houyoux et al., 2000). Emissions dependent on meteorological conditions are simulated during model execution through a CMAQ-pre-processor. Monthly mean lateral boundary conditions adopted from a species mapping methodology are incorporated into the models (Tang et al., 2008). A selected set of chemical fields derived from the GEOS-CHEM global model simulation with assimilated meteorology for July 2006 is employed (Bey et al., 2001).

The CB05 mechanism used in the NAQFC-β tends to produce 2.0 - 5.0 ppbv more ozone

Table 3.1: Locations and descriptions of the six surface monitors involved in the DISCOVER-AQ campaign.

Site Name	FIPS Code	Type	Lat [deg N]	Lon [deg E]	Elev. [m]
Aldino	240259001	Commercial Suburban	39.563	-76.204	127.711
Beltsville	240330030	Residential Suburban	39.028	-76.817	52.730
Edgewood	240251001	Military Rural	39.410	-76.297	8.534
Essex	240053001	Residential Suburban	39.311	-76.474	12.802
Fairhill	240150003	Residential Rural	39.701	-75.860	117.652
Padonia	240051007	Residential Suburban	39.461	-76.631	119.481

than the CB-IV mechanism used in the NAQFC (Sarwar et al., 2008) in the Mid-Atlantic region. Sensitivity tests on the CB05 mechanism found that additional reactions associated with organic peroxide species and reactive nitrogen recycling are primarily responsible for the additional ozone (Saylor and Stein, 2012). Although these results are useful, how these differences impact the utility of these models to the end user has been unknown until examined in detail here.

3.3 Data

Hourly average surface ozone forecasts for each day in July 2011 from both the NAQFC and the NAQFC-β were used in this statistical analysis. Forecasts were taken from the 1200 UTC model runs from the previous day since it is these forecast fields that air quality forecasters and decision makers use. The Maryland Department of the Environment (MDE) provided 1-h average near-surface ozone mixing ratios as measured by UV photometers at the six monitors of interest during the campaign (Table 3.1). These observations are used to verify the 1-h average surface ozone mixing ratio forecasts produced by both the NAQFC and the NAQFC-β from the previous day.

Figure 3.1 identifies the MDE monitor locations along with the nearest model pixels used in this analysis. The monitors generally lie near the periphery of the model grid points, with the exception of Fairhill, and the model grid points associated with Edgewood and Essex contain a significant amount of water from the Chesapeake Bay. Though the peripheral location of the monitors may induce small representation errors, the mixed land-water pixels at Edgewood and Essex may induce notable features in the performance metrics.

Decisions made in air quality management are generally associated with pollution exceedances with reference to the National Ambient Air Quality Standards (NAAQS). The primary NAAQS regarding surface ozone of 75 ppbv is based on a daily maximum of forward running 8-h average mixing ratios. Thus all ozone mixing ratios, both MDE observed and model predicted, have been averaged as such. For any given hour, the current 1-h average ozone mixing ratio was arithmetically averaged with the subsequent 7 hours of the 1-h average ozone mixing ratios.

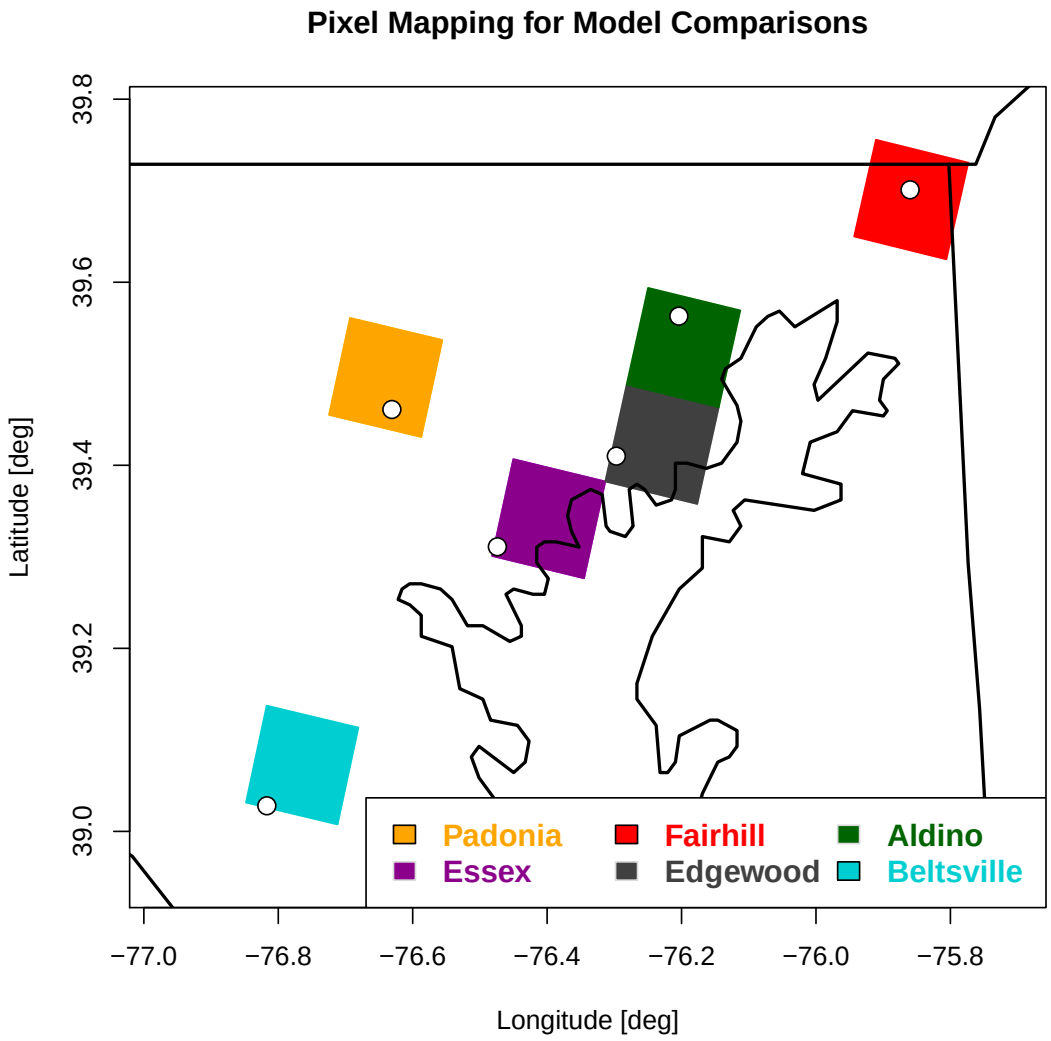

Figure 3.1: Map of the MDE monitor locations and model pixels used in the analysis. Pixels are color-coded according to the corresponding monitor. The monitor locations are indicated by the white dots within the model pixels.

Averages calculated with no more than 2 hours missing were included in the analysis.

3.4 Methods

3.4.1 Bootstrapping

A bootstrap sampling algorithm (Efron, 1979; Efron and Gong, 1983), or "bootstrapping", is used throughout this work to test for significant differences between the NAQFC and the NAQFC-β.

Bootstrapping involves repeatedly sampling a dataset with replacement in order to account for sampling uncertainties. The original dataset is considered a pool of samples from which bootstrap sub-samples are drawn. A statistical metric (i.e. bias, error, correlation, etc.), a model, or any other sort of analysis is then performed on the bootstrap sub-sample. Once the analysis is complete on this bootstrap sub-sample, another bootstrap sub-sample is drawn. This process continues for a user-defined (and often large) number of iterations.

Bootstrapping produces a distribution of results from the analyses performed on each bootstrap sub-sample. Various descriptive statistics (such as mean, median, quantiles, etc.) are then used to describe the distribution about the metric which in turn are analogous to the uncertainties about the metric of the original dataset. Often times, the 2.5% and 97.5% quantiles of the distribution are used to empirically represent the 95% confidence interval (CI) about the given metric. A quantity would then be considered significantly different from a given value if that value is not contained within the CI. Differences between two sources of data can be tested for significance by bootstrapping the same metric from each of the two sources and comparing the resulting CI from each distribution. Distributions of the same metric for different sources are considered significantly different if the CIs do not overlap.

To test for significant spatial and temporal differences between the NAQFC and NAQFC-β, bootstrapping was applied to each forecast hour in each pixel of the co-located model domain. For any given pixel and hour of the day, the 31 daily forecasts comprised the sample pool from which 10,000 bootstrap subsamples were produced. The CI about the mean residual, referenced to the NAQFC (i.e. NAQFC-β − NAQFC), were calculated from each bootstrap subsample. Insignificant differences will contain zero within the bounds of the CI. The minimum difference to insignificance (MDI) is defined as the CI bound closest to zero in those cases where the differences are significant or

$$MDI = \begin{cases} CI_{lo} & \text{if } CI_{lo} > 0, \\ CI_{up} & \text{if } CI_{up} < 0, \\ 0 & \text{if } 0 \in CI \end{cases} \tag{3.1}$$

where CI_{up} and CI_{lo} are the upper and lower bounds of CI respectively. The lower CI bound is used when the difference is positive (NAQFC-β > NAQFC) and the upper CI bound is used when the difference is negative (NAQFC-β < NAQFC).

Model bias and error were tested for significance at each site in Figure 3.1 (Table 3.1) using bootstrapping. For each hour of the day, the 31 available model forecasts and observations for that hour were used as the bootstrap sample pool. The mean bias (MB) and root-mean-squared-error (RMSE) were calculated for each of the 10,000 bootstrap subsamples taken from the sample pool. These metrics are often used to evaluate the performance of a continuous forecast system. Additionally, the difference between the NAQFC and NAQFC-β with respect to MB and RMSE were calculated to identify times of the day where the models significantly differ from one another.

Table 3.2: Contingency table for the simple decision scenario. F_i and O_i represent the forecasted and corresponding observed state of ozone.

	$F_i > \theta$ (Protect)	$F_i \leq \theta$ (Do not protect)
$O_i \leq \theta$ (Good AQ)	C	—
$O_i > \theta$ (Poor AQ)	C	L

3.4.2 Value of Information

A static cost-loss ratio model was used to quantify the value of information that forecast models provide (Thompson, 1952; Thompson and Brier, 1955; Katz and Murphy, 1997; Richardson, 2000). This statistical model directly relates various aspects of forecast skill to potential savings in expenditure. For this simple model, consider a decision scenario that requires an action based on the predicted mixing ratio of surface ozone. Let θ represent the threshold of surface ozone on which the action depends. Forecasted ozone greater than θ would constitute a forecasted ozone event and require a different action than when the forecasted ozone is less than θ. Let $\Theta = \mathbb{R}_{>\theta}$ represent all possible values of surface ozone greater than the defined threshold. The forecast hit rate (Hit), miss rate ($Miss$), and false alarm rate (FAR) are then defined as

$$Hit = \frac{\sum_{i=1}^{N} F_i \in \Theta \wedge O_i \in \Theta}{N} \tag{3.2a}$$

$$Miss = \frac{\sum_{i=1}^{N} F_i \notin \Theta \wedge O_i \in \Theta}{N} \tag{3.2b}$$

$$FAR = \frac{\sum_{i=1}^{N} F_i \in \Theta \wedge O_i \notin \Theta}{N} \tag{3.2c}$$

where F_i is the forecasted ozone corresponding to the observed value O_i and N represents the total number of forecast-observation pairs. These rates are often used to evaluate the performance of a binary forecast system by providing the frequency at which ozone events were forecasted and observed (Hit, Eq. 3.2a), observed but not forecasted ($Miss$, Eq. 3.2b), and forecasted but not observed (FAR, Eq. 3.2c).

Now, consider the contingency table (Table 3.2) that is associated with this simple decision model. One would take protective actions if $F_i > \theta$, or when the forecasted ozone is greater than a given threshold. This protective action comes at a cost C which fully insures against any loss L associated with an ozone event whether or not an ozone event occurs. If an ozone event is observed and no protective actions are taken, one would expect to lose L. Costs to protect are assumed to be less than the potential losses incurred, otherwise one would never protect and thus there would be no decision to make.

The value of a forecast system can then be defined as

$$V = \frac{min(\frac{C}{L}, s) - [(Hit + FAR)\frac{C}{L} + Miss]}{min(\frac{C}{L}, s) - s\frac{C}{L}} \tag{3.3}$$

where $s = Hit + Miss$ or the frequency of ozone events that exceed the threshold (Richardson, 2000; Garner and Thompson, 2012). The first term in the numerator and the denominator represents the expenditures due to the frequency of poor ozone events. The decision-maker would choose to either always protect or incur a loss proportional to the frequency of poor ozone events, whichever expenditure is the least. This provides the baseline from which the value of a forecast system is calculated. The second term in the numerator is the expenditure when using the forecast system. The decision-maker would spend C everytime the forecasted ozone is greater than the threshold and incur a loss of L everytime poor ozone occurred without being forecasted. The second term in the denominator represents the expenditure of a perfect forecast system. With a perfect forecast system, one would only spend C for each poor ozone event and nothing otherwise. Equation 3.3 has been normalized by L to cast value as a function of a cost-loss ratio.

From Equation 3.3, value can be interpreted as a savings over no forecast guidance relative to a perfect forecast. The cost-loss ratio can be interpreted as various protective measures from which the decision-maker may choose. With the bounds on C mentioned earlier, the cost-loss ratio will range from zero to one with large ratios corresponding to more expensive protective measures. Value is undefined for cost-loss ratios equal to zero or one. A perfect forecast system will have a value of one over all cost-loss ratios between zero and one. Typically, a forecast system will provide value between zero and one over a subset of cost-loss ratios.

To determine the relative value of the models, the static cost-loss ratio algorithm described above is used. The value is first calculated for each of the models separately. The number of hits, misses, and false alarms are counted with reference to the NAAQS for each model and used in Equation 3.3. The calculated value of the NAQFC, as a function of the cost-loss ratio, is then subtracted from the calculated value of the NAQFC-β to determine the relative difference in value between the two models.

3.5 Results and Discussion

3.5.1 Surface Evaluation

Figure 3.2 contains maps and histograms of the MDI between the NAQFC and NAQFC-β at three times of the day. The MDI between the two models is fairly homogeneous in the morning hours prior to maximum ozone production (Figure 3.2a,b). The distribution of the MDI is concentrated about the mean of 3.51 ppbv with an interquartile range of 0.97 ppbv. Large MDI at 0800 EDT (approximately 2-3 ppbv higher than the mean MDI) are observed in large cities, transportation routes, and elevated terrain along the Virginia/West Virginia border. Evidenced by the distribution of MDI and minimal emission impact, it is safe to interpret the background

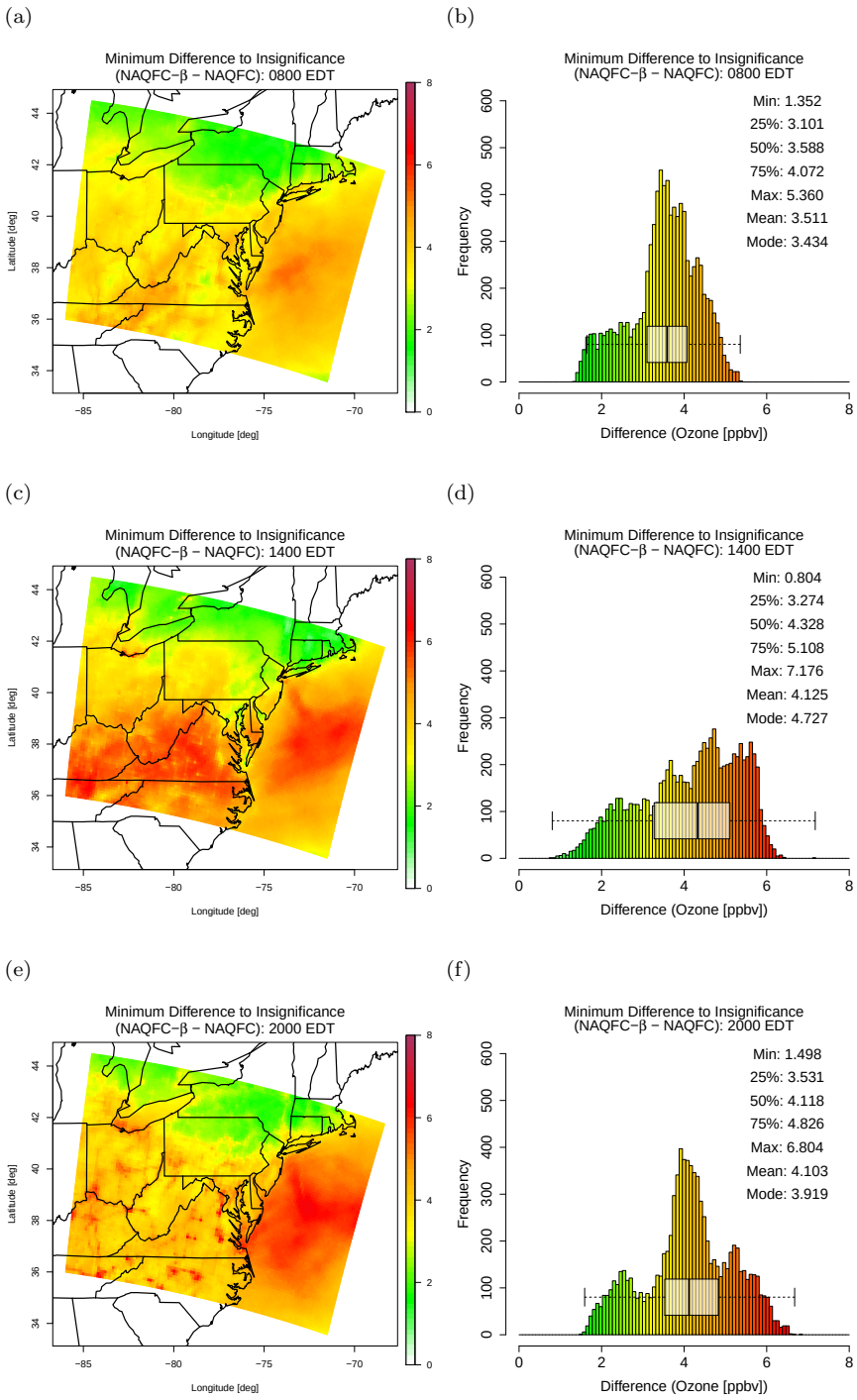

Figure 3.2: Maps of the MDI in forecasted surface ozone between the NAQFC and the NAQFC-β and corresponding histograms with boxplots at a,b) 0800 EDT, c,d) 1400 EDT, and e,f) 2000 EDT. A CI about the mean difference between the two models was calculated for each pixel. The color shading is MDI in ppbv and represents how far the derived CI bounds are from including zero. White pixels indicate insignificant differences. Shades of green, yellow, orange, and red indicate that the NAQFC-β is increasingly greater than the NAQFC.

Table 3.3: The statistics calculated from the plots in Figure 3.3 including the correlation (Corr.), root-mean-square error (RMSE), mean bias (MB), and the normalized mean bias (NMB) of the forecast models with surface observations. RMSE and MB are provided in units of ppbv. All statistics are significantly different from zero at the 95% CI. The relative difference between the NAQFC and the NAQFC-β with respect to each of these statistics are also significant with the exception of the correlations(indicated with *).

	Corr.		RMSE		MB		NMB	
	NAQFC	NAQFC-β	NAQFC	NAQFC-β	NAQFC	NAQFC-β	NAQFC	NAQFC-β
Aldino	0.70*	0.69*	15.56	16.39	-1.15	3.40	-2.28%	6.75%
Beltsville	0.82*	0.81*	16.86	20.12	8.96	13.84	22.34%	34.49%
Edgewood	0.69*	0.67*	18.75	20.92	4.47	9.03	9.75%	19.69%
Essex	0.71*	0.70*	18.29	20.66	5.65	10.39	12.88%	23.67%
Fairhill	0.73*	0.72*	13.59	15.81	3.54	7.81	7.98%	17.61%
Padonia	0.77*	0.77*	14.39	16.32	3.21	7.78	7.11%	17.22%

MDI between the two model versions as approximately 3.51 ppbv.

The distribution of MDI in the afternoon is more variable with an interquartile range of 1.83 ppbv (Figure 3.2c,d). Many of the features observed in the morning hours are no longer discernible. The elevated MDI appears to be contained within the southern portions of the domain and over the Atlantic Ocean. At 2000 EDT (Figure 3.2e,f), the diffuse regional MDI over the land mass is replaced by local maxima collocated with emission sources and transportation routes. The ambient MDI drops back to the background levels similar to those observed in Figure 3.2a. Transportation routes, such as interstates, state highways, and waterways, exhibit MDI of approximately 5 ppbv. Large cities are associated with the largest MDI, reaching over 6 ppbv.

The differences observed in the MDI are likely due to the advanced recycling methods of reactive nitrogen in the CB-05 mechanism present in the NAQFC-β. The reactive nitrogen recycling effectively converts stable forms of nitrogen into more reactive forms. Incorporating these reactions into the model produce additional pathways through which ozone can be produced, ultimately increasing ozone concentrations.

Figure 3.3 and Table 3.3 describe the general skill of the NAQFC and NAQFC-β in forecasting 1-h average surface ozone over the six DISCOVER-AQ sites. The median of the forecasts indicate that overprediction at the lowest observed surface ozone mixing ratios (0 - 30 ppbv) is common among all six of the sites. The forecasts tend to follow the 1:1 line between 30 - 75 ppbv of observed ozone mixing ratios. There are too few observation-forecast pairs to definitively say that the models underpredict at high levels of observed ozone ($>$ 75 ppbv), but the plots in Figure 3.3 show that this trend is plausible among many of the sites. These descriptive statistics of the comparisons indicate that the NAQFC outperforms the NAQFC-β. The differences between the statistics reported for each model were found to be significant with the exceptions being the correlations.

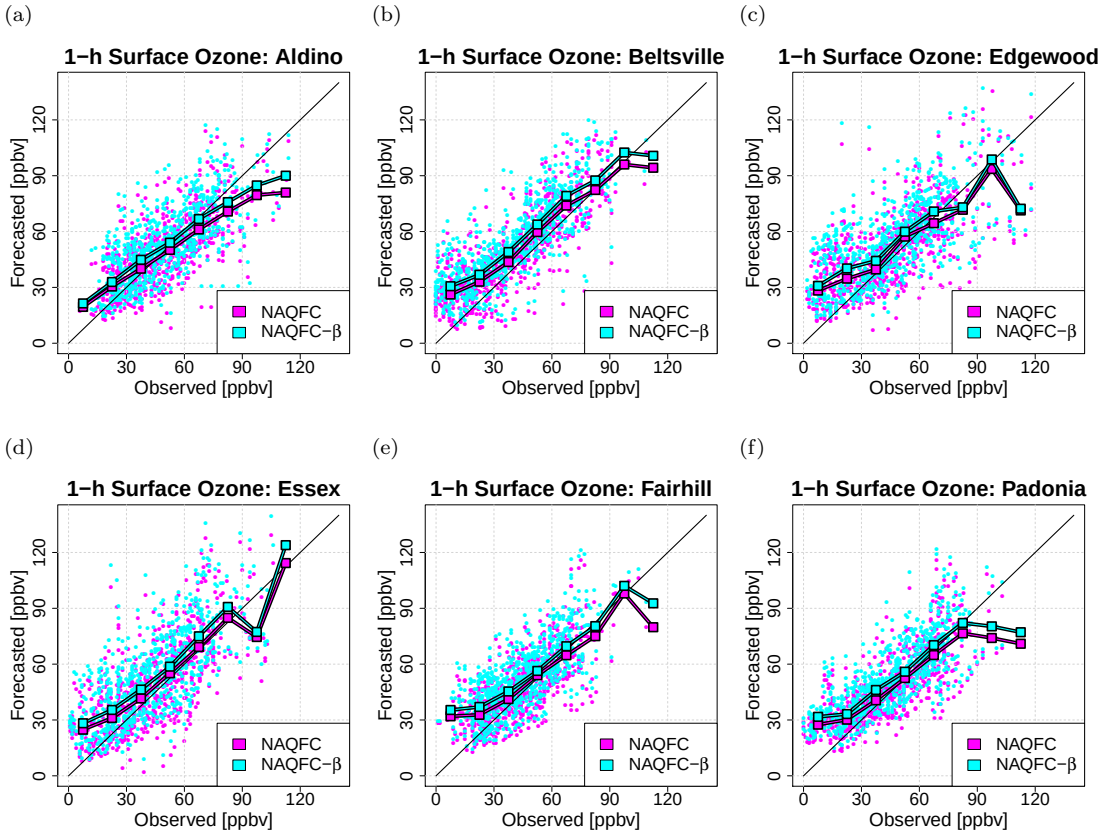

Figure 3.3: Summary of the general skill of the NAQFC (magenta) and the NAQFC-β (cyan) in forecasting 1-h average surface ozone at a) Aldino, b) Beltsville, c) Edgewood, d)Essex, e) Fairhill, and f) Padonia. Each dot represents a single observation-forecast pair. The squares indicate the median forecast for a 15 ppbv bin of observed ozone. The 1:1 line is provided for guidance.

The bias among all the sites follow a similar diurnal pattern (Figure 3.4). Generally, there is a positive mean bias in the early morning hours that decreases throughout the morning into the evening. This is consistent with the overprediction of low ozone values evidenced in Figure 3.3. At 1900 EDT, approximately an hour before sunset, the mean bias begins to increase again to the levels found in the early morning. The exception to this pattern is Beltsville (Figure 3.4b) where the bias remains fairly consistent throughout the day.

The extent to which these biases are significant varies from site to site. The bias at Beltsville for both models is significant for 23 hours of the day, losing significance at 2100 EDT. Aldino (Figure 3.4a) is the only site to display a significant negative bias for both models, occurring between 1800 EDT and 1900 EDT. The diurnal pattern at Fairhill and Padonia (Figures 3.4e,f respectively) closely resembles that of Aldino, only positively shifted by approximately 7 ppbv and 10 ppbv, respectively, throughout the morning and afternoon. This yields significant positive differences between 0000 EDT 1100 EDT at both locations. Both Edgewood and Essex

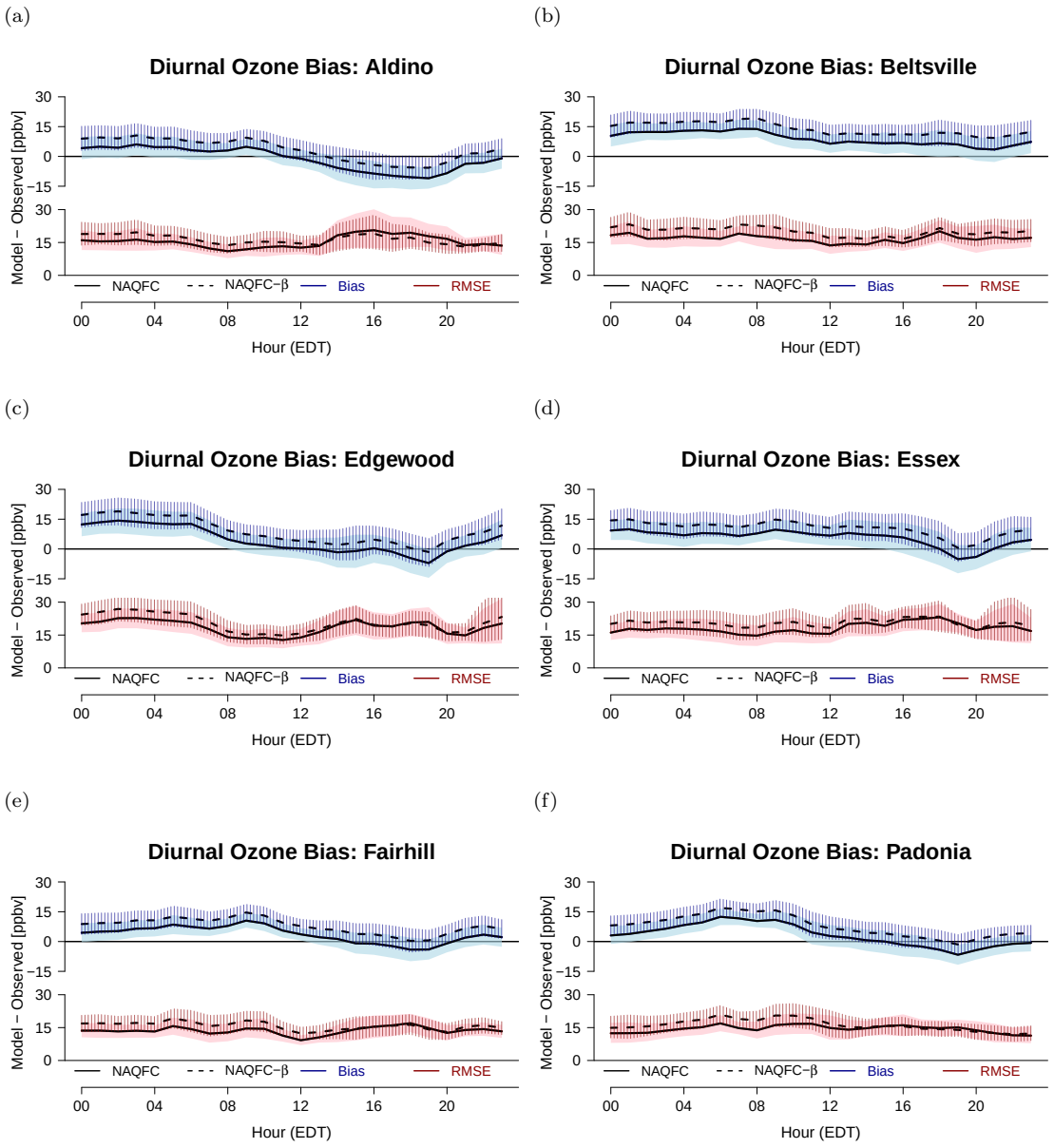

Figure 3.4: Bias and RMSE as a function of the hour of the day for a) Aldino, b) Beltsville, c) Edgewood, d)Essex, e) Fairhill, and f) Padonia. The model type is denoted by line and fill type (NAQFC - solid; NAQFC-β - dashed). The statistic is color-coded (Bias - blue; RMSE - red). The mean and CI of each statistic is indicated by the line and fill respectively.

(Figure 3.4c,d) exhibit a sharp decrease in bias at 1900 EDT followed by a sharp increase the following hour. The proximity of these two sites to the bay exposes the monitors to bay-breeze circulations which are known to cause a spike in ozone at approximately this time (Stauffer et al., 2012).

The RMSE exhibits few common features among the sites. The average RMSE tends to be greater in the afternoon (approximately 22 ppbv) compared to the morning (approximately 15 ppbv) with the exception of Beltsville, Padonia, and (to a lesser extent) Fairhill with fairly constant RMSE throughout the day. The variability in the RMSE for any given hour, evidenced by the CI, is small at Fairhill and Padonia relative to the other sites with CI of approximately 7 - 10 ppbv. In contrast, the greatest variability for any given hour is found at Edgewood and Essex with CI of approximately 10 - 20 ppbv. This hourly variability at Aldino is low throughout most of the day except in the afternoon (1400 - 2000 EDT) where the CI more than doubles from 7 ppbv to 15 ppbv. Similar spikes in CI occur in the NAQFC-β at Edgewood and Essex after 2100 EDT. These spikes seem to coincide with periods during or shortly after large changes in the bias, though this is not true for relatively smaller spikes at other sites. This is likely due to the models incorrectly forecasting these late-day transition periods.

Figure 3.5a depicts the differences in the bias between the NAQFC and NAQFC-β as a function of hour of the day. The NAQFC-β consistently forecasts 1-h surface ozone 4 ppbv greater than the NAQFC throughout the day. There are slight inflections in the bias difference at all the sites at 0900 EDT, 1300 EDT, and 2100 EDT. These inflections are most pronounced at Beltsville, Edgewood, and Essex; however these inflections constitute changes in the bias on orders of less than 1 ppbv.

Figure 3.5b is the difference in RMSE between the two models at the six sites as a function of hour of the day. There is more diurnal variability in the RMSE differences when compared to the bias. Generally, the difference in RMSE between the two models decreases throughout the day, minimizing in the late afternoon, and rising back to pre-dawn levels after 2000 EDT. The exception is Padonia which starts the day low and peaks at 0900 EDT before becoming similar to the other sites. The CI generally increases into the afternoon hours at all sites except Aldino which stays fairly constant throughout the day.

3.5.2 Value of Information

The maximum forecasted 8-h average ozone does not always coincide with when the maximum is actually observed. To analyze this, a scatterplot of the hours at which the maximum 8-h average ozone was forecasted and observed is shown in Figure 3.6. Histograms of the difference between the forecasted and observed hour of maximum 8-h average ozone are provided in the lower-right of each plot. Both models produced the same hours of maximum ozone for each day at all sites, so the forecasted hour in Figure 3.6 refers to both the NAQFC and NAQFC-β. Generally, the models do well at forecasting the hour of maximum 8-h average ozone. The models forecast the hour correctly between 16.7% (Essex) and 40.0% (Fairhill and Padonia) of the time while

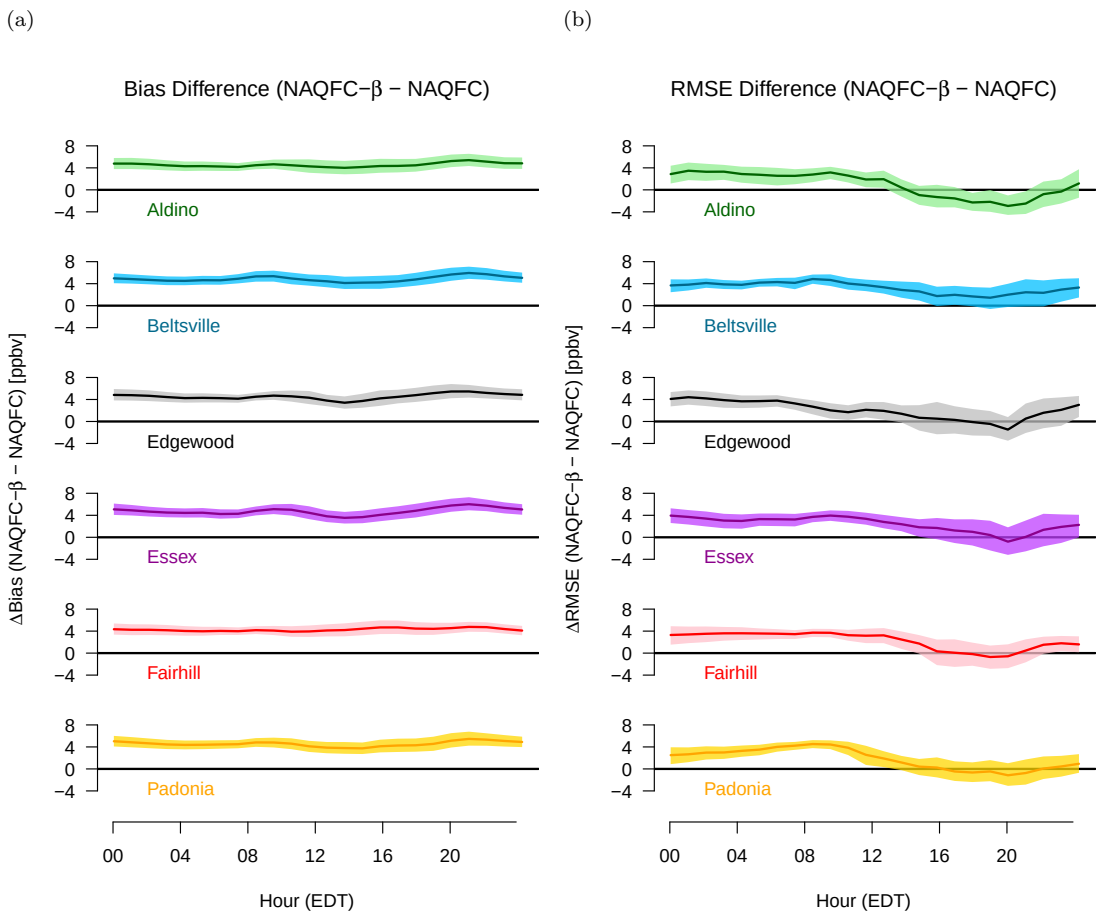

Figure 3.5: The difference in a)the bias and b) the RMSE between the NAQFC and NAQFC-β using the NAQFC as the reference value (NAQFC-β−NAQFC).

forecasting within one hour of the maximum between 56.7% (Aldino) and 73.3% (Beltsville) of the time. The distribution of the difference in forecasted and observed hours of maximum 8-h average ozone is biased slightly low with means ranging from -0.17 (Beltsville) to -1.10 (Aldino) and medians of -0.5 (Aldino and Essex) or zero. This is to be expected because the 1-h forecasted ozone at all sites starts off with a high bias in the late morning that decreases into the afternoon (Figure 3.4). When producing the running average, the hour at which the maximum occurs in the model will be skewed towards the time with the greater bias relative to the observations.

There is one outlier at Aldino, Edgewood, and Fairhill where the forecasted hour of maximum 8-h average ozone is at hour 0000 EDT. These points are all associated with 08 July. The model runs erroneously put a large cloud fraction over the northern half of the DISCOVER-AQ campaign area (not shown), keeping afternoon ozone production low and thus yielding a maximum 8-h average ozone that occurs at the start of the day. Scattered afternoon cloud cover allowed enough actinic flux to produce an ozone peak before becoming overcast and precipitating.

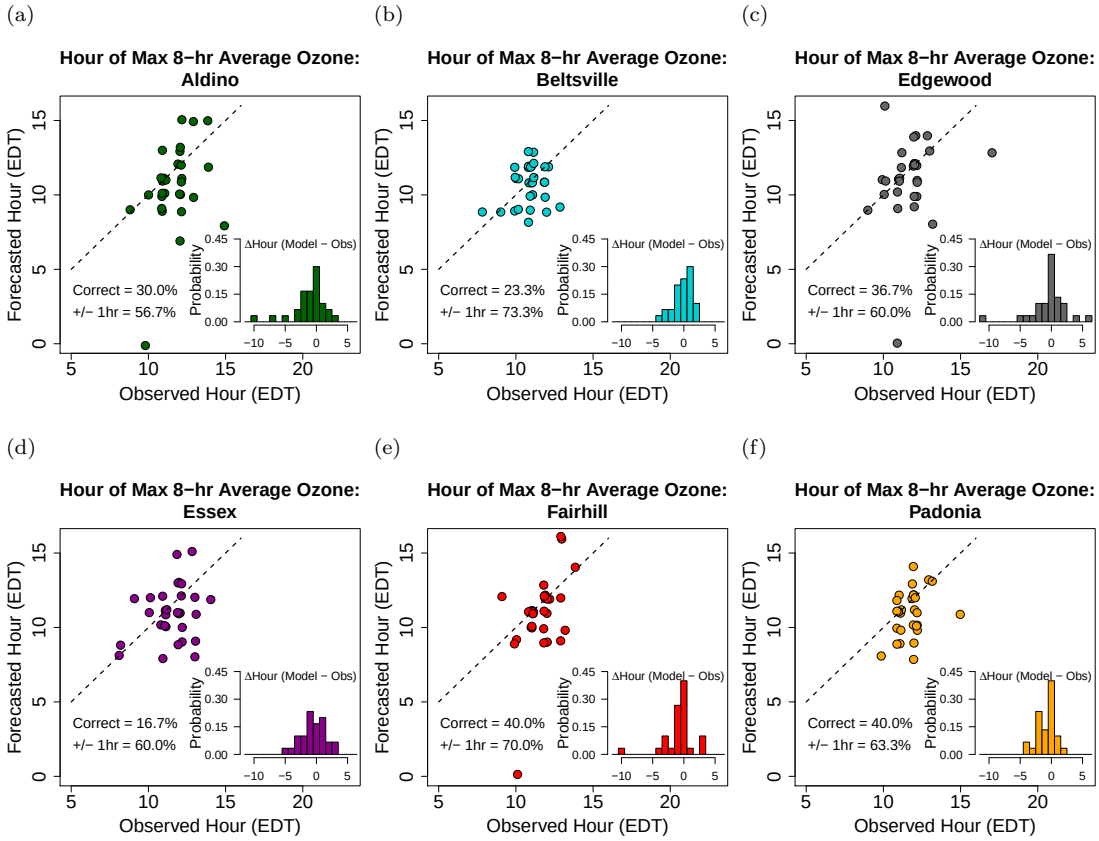

Figure 3.6: Scatterplot of the hour (EDT) at which the maximum 8-h average ozone was forecasted versus when it was observed at a) Aldino, b) Beltsville, c) Edgewood, d) Essex, e) Fairhill, and f) Padonia. The points are jittered slightly so that multiple points at the same coordinates are easily viewable. A dashed 1:1 line is provided for clarity. A histogram depicts the frequency of the difference between the forecasted and observed hour of maximum 8-h average ozone expressed as a probability.

Scatterplots of the daily maximum 8-h average ozone are shown in Figure 3.7. The hit rates, miss rates, and false alarm rates are provided in their respective quadrants in each plot. The differences in these rates between the two models exhibit the behavior one would expect when analyzing two similar yet unequally biased models. The NAQFC-β, which was shown in Figure 3.5a to have a significant positive difference in bias relative to the NAQFC, generally produces more hits and false alarms while reducing misses. In Beltsville, where the NAQFC fails to register a missed forecast, the higher biased NAQFC-β registers only an increase in false alarms.

The small differences in the hits, misses, and false alarms between the two models have large impacts on the value of information to the end user. Figure 3.8 is the difference in value of information between the NAQFC and NAQFC-β as a function of cost-loss ratio. The higher biased NAQFC-β, relative to the NAQFC, tends to convert "missed" forecasts into "hit" forecasts.

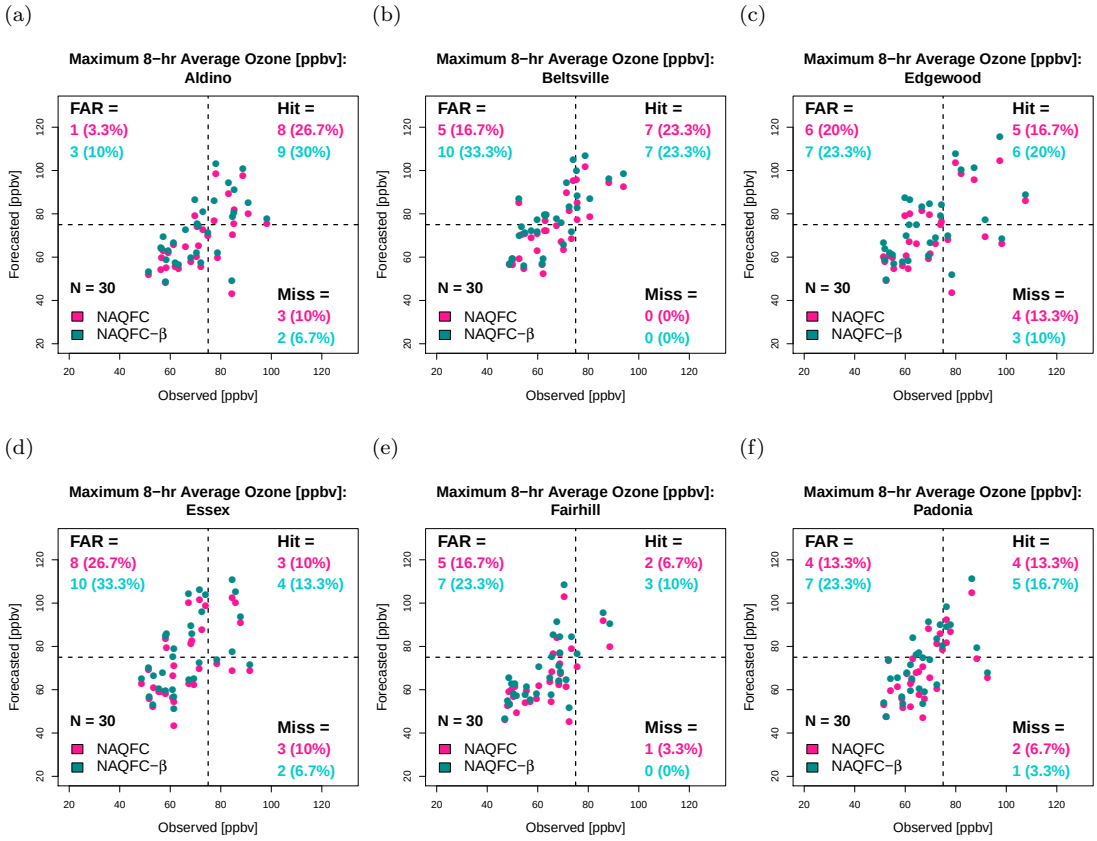

Figure 3.7: Summary of the general skill of the NAQFC (magenta) and the NAQFC-β (cyan) in forecasting maximum daily 8-h average surface ozone at a) Aldino, b) Beltsville, c) Edgewood, d)Essex, e) Fairhill, and f) Padonia. The dashed lines indicate the current NAAQS standard of 75 ppbv for an 8-h average ozone mixing ratio. The false alarm rate (FAR), hit rate (Hit), and miss rate (Miss) are provided as the number of observation-forecast pairs and corresponding percentage in parentheses.

This effectively increases the maximum relative value of the NAQFC-β while simultaneously reducing the minimum cost-loss ratio at which the NAQFC-β produce value. Coincidently, the increased false alarm rate tends to reduce the maximum cost-loss ratio at which the NAQFC-β produces value (Wandishin and Brooks, 2002).

The most dramatic differences are at Aldino, Beltsville, Fairhill, and Padonia. The differences in value at Aldino and Padonia switch from positive at low cost-loss ratios to negative at high cost-loss ratios. This indicates that both the NAQFC and the NAQFC-β must be used together at these two sites to optimize the value over a broad range of decision scenarios. The NAQFC provides greater forecast value at Beltsville through a cost-loss ratio of 0.61 at which point both models produce equal value. The NAQFC-β produces greater forecast value at Fairhill through a cost-loss ratio of 0.30 before the difference in value between the models becomes negligible. The NAQFC-β provides slightly greater forecast value than the NAQFC at the Edgewood and Essex

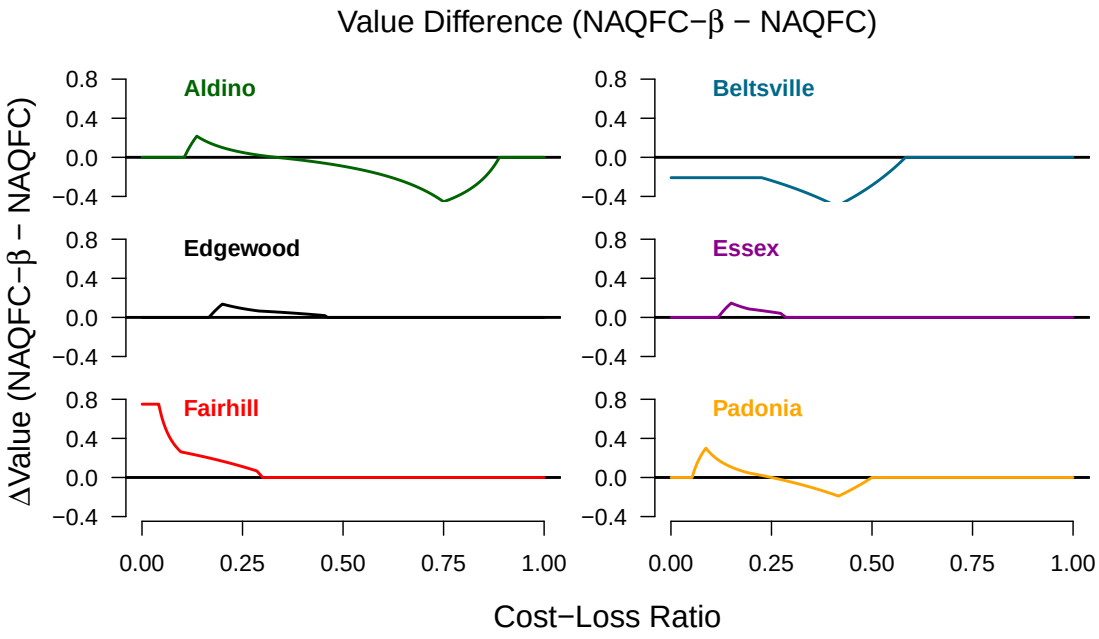

Figure 3.8: Relative difference in value between the NAQFC and the NAQFC-β as a function of the cost-loss ratio.

sites, though this difference is small and over a short range of cost-loss ratios.

3.6 Summary and Conclusions

Two numerical air quality models provided forecast support in flight decisions during the 2011 DISCOVER-AQ campaign. The updated chemical mechanism in the NAQFC-β tends to produce higher surface ozone mixing ratios than the mechanism used in the NAQFC (Saylor and Stein, 2012). Statistical tests were performed to evaluate the skill of these two numerical models in predicting surface ozone, to determine the statistical significance of any differences in surface ozone predicted by the two models, and to assess the change in the value of information as a result of the updated mechanism.

A domain-wide analysis of the differences in 24-h forecasted surface ozone revealed significant differences between the two models. The MDI indicates that the NAQFC-β is at least 3.51 ppbv higher than the NAQFC. These background differences are fairly homogeneous throughout the domain during the early morning hours with a few maxima associated with the locations of known emissions sources standing out as slightly greater than the mean. Both the mean and the spread of these differences increase in the afternoon primarily in the southern portions of the model domain and over the Atlantic ocean. The regional nature of the elevated MDI fades into the evening hours revealing MDI maxima up to 6 ppbv located over emissions sources. These results

confirm that the CB05 chemical mechanism in the NAQFC-β produces statistically significant differences in the predicted surface ozone compared to the NAQFC.

The skill of these two models was analyzed at six surface sites of interest in the DISCOVER-AQ campaign. The standard descriptive statistics indicate that the NAQFC outperforms the NAQFC-β. The correlation between the NAQFC and surface observations were either identical or slightly better than the correlation between the NAQFC-β and surface observations. The NAQFC also provided predictions with significantly less bias and significantly less error than the NAQFC-β at all six sites. Both models typically overpredict surface ozone in low ozone regimes (0 - 30 ppbv) while tending to underpredict in high ozone regimes (75 - 90 ppbv).

The bias diurnal pattern for both models is consistent throughout the day among all of the sites. The bias is significantly high in the morning hours up until the hours of maximum ozone production at which time the biases typically drop to insignificant levels. The only exceptions to this are Beltsville, where the bias is significantly high throughout the day, and Aldino, which is the only site in which the bias becomes significantly negative in the afternoon. The CIs about the mean biases are fairly stable throughout the day, only slightly increasing in the afternoon hours at all the sites.

The RMSE diurnal patterns have few common features among all the sites. The RMSE typically decreases throughout the morning hours until approximately 1300 EDT when the RMSE begins to rise. Beltsville, like in the bias diurnal patterns, is fairly constant throughout the day. Padonia exhibits a fairly constant RMSE diurnal profile with the largest variability occurring in the morning according to the CI. The CI at the other sites tends to start off small and increase in the afternoon and evening hours. This increase in variability in RMSE can be sharp at sites including Aldino, Edgewood, and Essex.

The bias is significantly higher in the NAQFC-β than the NAQFC throughout the entire day. Slight perturbations in the diurnal pattern occur at the ozone transition times, though these account for less than a 1 ppbv variation in the difference in the bias. The differences in the RMSE tend to start high in the morning hours and decrease to insignificant differences in the afternoon. The only exception is Beltsville where the NAQFC-β remains significantly high all day.

The descriptive statistics indicate that the NAQFC performs significantly better than the NAQFC-β in predicting surface ozone mixing ratios; however, the utility of the model is dependent on the needs of the end user. A static cost-loss ratio model was used to assess the relative difference in the value each of these models provide the user. The NAQFC-β produced greater value of information, typically at low cost-loss ratios, relative to the NAQFC. Beltsville was the only site in this analysis where the NAQFC provides more value than the NAQFC-β. Aldino and Padonia are the only sites where a combination of both models would yield the best overall value in decision-making. This is counter-intuitive, but the less skillful model produces greater value of information than the more skillful model for certain decisions. This is because standard evaluation metrics often mask the sensitivity of the end users' needs to forecast error.

Chapter 4

Ensemble Statistical Post-Processing of the National Air Quality Forecast Capability: Enhancing the Value of Ozone Forecasts in Baltimore, Maryland

4.1 Introduction

Baltimore, MD is ranked the ninth most ozone polluted metropolitan area in the U.S. according to the 2013 State of the Air released by the American Lung Association (`http://www.stateoftheair.org/2013`), sharing the top spots with cities such as Los Angeles, CA and Houston, TX. Consequently, forecasting ozone is vital to the health and well being of over eight million Baltimore residents. Expert air quality forecasters use a combination of numerical models, statistical models, and empirical rules to provide accurate and timely forecasts. These tools generally provide skillful (Eder et al., 2010) and valuable (Garner and Thompson, 2012) forecasts; however, these tools have limitations that pose a unique forecast challenge in Baltimore.

The Baltimore forecast region is depicted in Fig. 4.1. The region encompasses counties of Maryland that are located to the north and west of the Chesapeake Bay. Eight ozone monitors, described in Table 4.1, comprise the monitoring network within the Baltimore forecast region. Six of the eight monitors lie within 30 km of the bay.

Near-surface ozone is produced through photochemical reactions with nitrogen oxides (NO_x) and volatile organic compounds (VOCs) (Seinfeld and Pandis, 2006); thus seasonal ozone concentrations peak during the late-spring through early-fall months (April - October; Ozone Season)

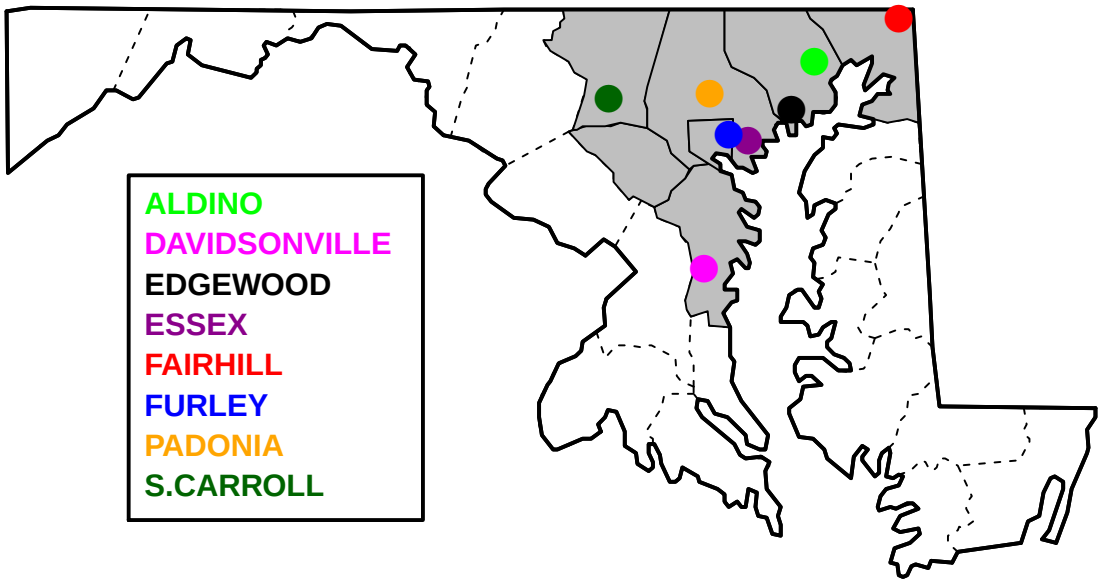

Figure 4.1: Map of ozone monitors in the Baltimore, MD forecast region. The forecast region is shaded in gray according to the region definition provided by the Maryland Department of the Environment.

Table 4.1: Baltimore, MD air quality monitor locations.

Site Name	FIPS Code	Lat [deg N]	Lon [deg E]	Elev. [m]
Aldino	24-025-9001	39.563	-76.204	127.7
Davidsonville	24-003-0014	38.903	-76.653	44.0
Edgewood	24-025-1001	39.410	-76.297	8.5
Essex	24-005-3001	39.311	-76.474	12.8
Fairhill	24-015-0003	39.701	-75.860	117.7
Furley	24-510-0054	39.329	-76.553	49.0
Padonia	24-005-1007	39.461	-76.631	119.5
South Carrol	24-013-0001	39.444	-77.042	226.0

when ample solar radiation is available for the photochemical reactions. Daily ozone concentrations, however, are strongly driven by the local meteorology (Dimitriades, 1976). Cloud-free skies maximize the actinic flux available to the photochemical reactions while light winds promote stagnation and aggregation of ozone and its precursors. These meteorological features also create an ideal environment through which a bay breeze may form within the Baltimore forecast region (Banta et al., 2005). The temperature gradient between the warm solar-heated land and

the cool water creates a thermal circulation which can concentrate ozone and its precursors along the airmass boundary (Stauffer et al., 2012; Stauffer and Thompson, 2013). The current suite of regional operational numerical models are run at resolutions that are too coarse to forecast the onset and location of bay breeze events to the degree needed for assessing the impacts on local air quality (Banta et al., 2005). This includes the National Air Quality Forecast Capability (NAQFC), the current national air quality model produced by the National Oceanic and Atmospheric Administration (NOAA) and the Environmental Protection Agency (EPA) (Janjic, 2003; Byun and Schere, 2006; Garner et al., 2013), with an operational horizontal resolution of 12 km. Loughner et al. (2011) found that a horizontal resolution of 4.5 km or finer in numerical meteorological models produced discernible simulations of the bay breeze in the Baltimore region. Without properly resolving the bay breeze, the NAQFC will not be able to properly handle ozone predictions along coastal boundaries such as those in the Baltimore forecast region.

Attaining uncertainty about the forecast from a numerical model is difficult. Common practice is to run an ensemble of numerical models, each with slightly perturbed initial conditions, boundary conditions, and/or parameterizations such as the NOAA Short Range Ensemble Forecast (SREF; `http://www.spc.noaa.gov/exper/sref/fplumes/`). From the suite of models used in the ensemble, one can determine a consensus prediction and spread from which uncertainties are derived. In order to reduce the computational burden, often times the members within the ensemble are run at a reduced resolution. This process is not feasible for the NAQFC for reasons described earlier. Alternatively, air quality forecasters are using multiple numerical models from different sources to create a "poor-man's" ensemble (Djalalova et al., 2010). Though this method was shown to improve upon the forecast from any single ensemble member, the ensemble would rely heavily on the individual model providers to continue producing the forecasts. If a single model provider decides to terminate their model, due to lack of funding for example, then the entire ensemble suffers.

In addition to numerical models, air quality forecasters often use statistical models derived from local data. These statistical models range from simple regression models to complex artificial neural networks (Thompson et al., 2001; Al-Alawi et al., 2008; Pires and Martins, 2011). The skewed distribution of ozone, as evidenced in Fig. 4.2, means that many of the standard statistical approaches may not be valid and that methods that incorporate extreme-value theory are preferred (Thompson et al., 2001).

The goal of this article is to develop a tool that can address these problems while providing the best value of information to the end user. The product must adjust appropriately for various local meteorological regimes and provide uncertainty information about the forecast while avoiding the pitfalls of forecasting extreme values. An ensemble statistical post-processor (ESP) for the NAQFC is a logical choice for such a product.

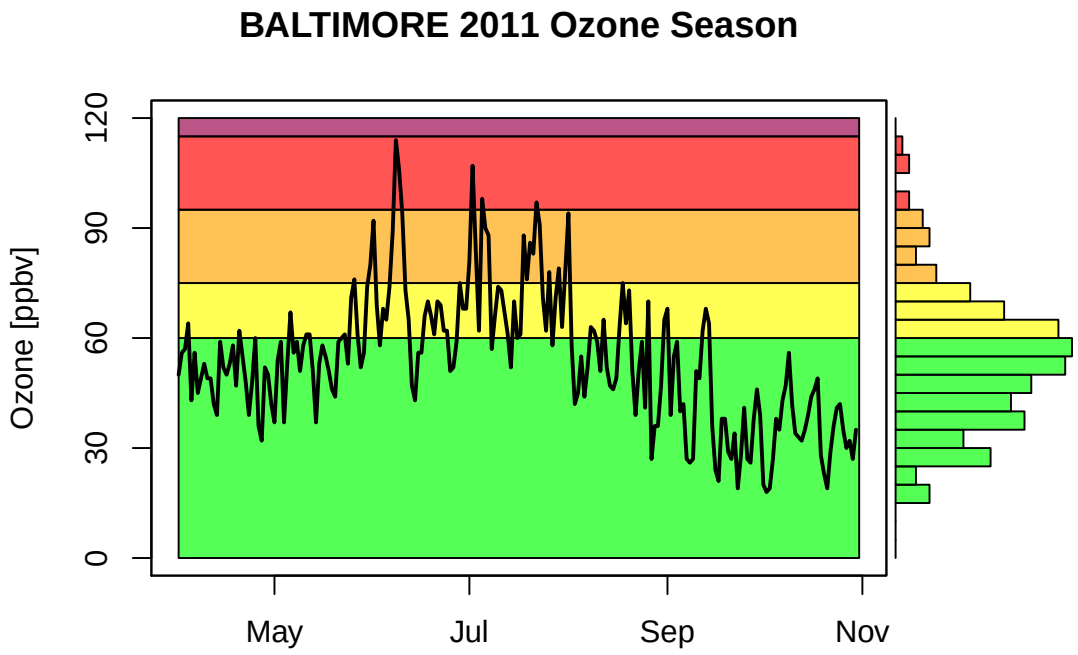

Figure 4.2: Time-series of the 2011 daily maximum 8-hr average ozone in Baltimore, MD. The ozone-season is defined as 01 April through 31 October. The histogram along the right margin is positively skewed suggesting that statistical models built on assumptions of normality using these data may result in underforecasting the ozone exceedance events. The background is shaded according to the air quality index.

4.2 Data and Methods

Data were collected for eight air quality monitoring locations in the Baltimore, MD forecast region shown in Fig. 4.1 and described in table 4.1. Hourly ozone observations (parameter code 44201) from the 2005 - 2011 ozone seasons (April - October) were collected from the EPA Technology Transfer Network (TTN) Air Quality System (AQS), a quality-controlled national database of atmospheric particle and trace-gas measurements (`http://www.epa.gov/ttn/airs/airsaqs/detaildata/downloadaqsdata.htm`). The National Ambient Air Quality Standard (NAAQS) for ozone is calculated using forward-running 8-h average concentrations, so the hourly ozone data were averaged as such. For any given hour, the ozone concentration for that hour is averaged with the subsequent seven hours of ozone data. This average represents the 8-h average for that particular hour. This averaging is performed for each hour in the data set. The daily maximum 8-h averages are used for the product development and validation. An ozone exceedance is defined in the NAAQS as a day with a maximum 8-h average ozone concentration greater than 75 ppbv.

The daily 1200 UTC NAQFC model output was collected from the NOAA National Operational Model Archive and Distribution System for the same dates as the hourly ozone obser-

Table 4.2: Meteorological variables used in the development of the ESP product.

Variable	Time/Type	Units
Temperature	Max, Min, 1800 UTC	K
Dewpoint Temperature	Max, Min, 1800 UTC	K
Sea-level Pressure	Max, Min	hPa
Relative Humidity	Max, Min	%
Sky Cover	1200 UTC, 1800 UTC	%
U-component Wind	1200 UTC, 1800 UTC	ms^{-1}
V-component Wind	1200 UTC, 1800 UTC	ms^{-1}
Precipitation	24-h Total	mm
Bay Breeze Index	1800 UTC	—

vations. The NAQFC model output includes the 8-h running average ozone, so no additional averaging is necessary. The maximum 8-h average ozone forecasted for the day following the initial model run date is used.

Historical meteorological observations collocated with the ozone monitors were collected from the National Climatic Data Center. In the event that meteorological information is not available for an ozone monitor, meteorological data from surrounding sites were interpolated to the monitor using a kriging algorithm (Ribeiro, Jr. and Diggle, 2001). The meteorological data set is listed in Table 4.2. The bay breeze index is a derived quantity (Sikora et al., 2010) that represents the degree to which the atmosphere favors bay breeze formation. The index is calculated relative to meteorological data from the buoy station TPLM2 at Thomas Point, MD. Including such an index will help the model development process sort out days with possible bay breeze events.

4.2.1 ESP Development

The ESP development mimics that of a perfect prog system (Klein et al., 1959; Wilks, 2011). Observational data are used to train the ESP. Forecasted quantities of the variables used in development are then used in the ESP to predict the future ozone concentrations. This method is preferred over model output statistics (Glahn and Lowry, 1972) when forecasting air quality due to the young age and rapid development of the NAQFC. As the NAQFC matures, the increase in forecast skill would translate over into the ESP.

This development process is applied to each monitor separately. The meteorological variables in Table 4.2 along with the NAQFC data are used as independent variables to predict the dependent variable ozone. The ozone concentration from the previous day is also included as an independent variable in order to take advantage of any serial correlation in the ozone data. These data are resampled 100 times with replacement using a moving-block bootstrap algorithm to produce 100 bootstrap subsamples of equal length of the original data set (Efron, 1979; Efron and Tibshirani, 1993; Wilks, 2011).

Each bootstrap subsample is fit to a unique regression tree model (Breiman, 1984). The

purpose of a regression tree model is to recursively split the ozone into homogeneous groups called nodes using the independent data. Splits and terminal nodes are typically scored with a metric based on the standard deviation of the groups, but in order to better predict an ozone exceedance, the f-measure is used instead (Torgo and Ribeiro, 2003; Ribeiro and Torgo, 2006). The f-measure

$$F = \frac{(\beta^2 + 1) \cdot precision \cdot recall}{\beta^2 \cdot precision \cdot recall} \tag{4.1}$$

is rooted in extreme-value theory and used to predict outliers of a data set by accounting not only for the homogeneity of the group, but also the ability of the group to recall the extreme values. The variable *precision* is $1 - NMSE$ or a function of the normalized mean square error of the predictions for extreme values. The variable *recall* is the ratio of predicted extreme values to the number of extreme values in the data set. β is a parameter that adjusts the relative importance of *precision* to *recall*. By defining an extreme value of ozone as the NAAQS of 75 ppbv and using the f-measure as a split function in these regression trees, the terminal nodes will be tailored to ozone exceedances. The path to the terminal nodes will represent local meteorological regimes conducive to producing ozone exceedances. The terminal nodes contain homogeneous clusters of ozone and the independent data describing it. Multivariate linear regression models are then fit to the data in the terminal nodes.

The resulting 100 regression tree models constitute the ESP for the given monitor. Three parameters are used in the development process. First, the minimum number of data instances for a given node was set to 30. This ensures enough data in a terminal node from which a regression model can be fit. The second and third parameters are the β parameter in the f-measure and the node-termination threshold of the f-measure (f-limit). Each was set to 0.8. These values were determined using a simple optimization scheme.

4.2.2 Parameter Optimization

A simple optimization scheme was used to estimate the β and f-limit parameters in the ESP development. In both Torgo and Ribeiro (2003) and Ribeiro and Torgo (2006), these parameters are set with precision of a single decimal point. Since both these values range from zero to one, all combinations of β and f-limit values were tested. The f-limits of zero and one were not included in the optimization process. Using an f-limit of zero would prevent the regression tree from splitting beyond the root node while using an f-limit of one would force the tree to grow until each node contains the defined minimum number of instances. This results in 99 unique combinations of β and f-limits (11 values of β and nine values of f-limit).

To perform the optimization, 10% of the training data set is reserved to evaluate the selected parameters. The remaining 90% is used in the training process described in section 4.2.1. Each combination of parameters was tested using the area under the relative operating characteristic (ROC) curve (see section 4.3 and Fig. 4.6 for additional details on the ROC curve). The combination that produces the maximum area under the ROC curve would be the ideal combination of parameters to use in the final ESP product.

A single set of parameters was used to develop all of the ESP products. Fig. 4.3 shows the results of the optimization process. Holding a β value constant, the ROC areas for each given value of the f-limit were bootstrapped into 10,000 sub-samples. The ROC areas in each sub-sample were centered on the median of the sub-sample, producing the deviations from the median shown in Fig. 4.3. For any given β value, the greatest positive deviations are associated with f-limits above 0.7. Essentially, any f-limit greater than 0.7 should result in high ROC area.

The same process is applied to the β parameter and shown in Fig. 4.4. Note that the deviations in the median are much smaller than those found in determining an f-limit, most by an order of magnitude. This indicates that the choice of β value will have a small impact on the overall performance and value of the ESP as long as the f-limit is set to a value above 0.7. A value of 0.8 was chosen for both the β and f-limit parameters because not only is it greater than 0.7 for the f-limit, but it's the only combination that produces the maximum deviations in both Fig. 4.3 and Fig. 4.4.

4.2.3 Cross-Validation

The ESP product was evaluated using a 10-fold cross-validation scheme (Picard and Cook, 1984; Wilks, 2011). The full data set is split into 10 groups each containing 10% of the original data set. Nine of the groups are used in the development process described in section 4.2.1 while the tenth group is reserved for evaluation. This process is repeated until each of the 10 groups is used as a reserved evaluation data set. Cross-validation ensures that the product is never evaluated with the same data used to build the product, resulting in evaluation metrics that closely represent skill in true operational forecasts.

4.3 Results

Probabilities are derived from the ensemble predictions. The number of predictions above the ozone standard of 75 ppbv is divided by the number of available ensemble predictions. This results in the forecasted probability of the monitor exceeding the ozone standard for the day. Baltimore regional probability forecasts were calculated from the individual monitor probability forecasts through a recursive application of the additive law of probability for non-mutually exclusive events

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{4.2}$$

This expression produces the forecasted probability of any given site within the Baltimore forecast region exceeding the ozone standard. Results of the ESP cross-validation for the Baltimore forecast region are shared here.

The attributes diagram in Fig. 4.5 describes the full joint distribution of the ESP forecasts and observed ozone in the Baltimore region (Hsu and Murphy, 1986). The observed relative frequency of ozone exceedances is plotted as a function of the ESP forecasted probability. The diagram would lie on the 1:1 line when using a perfect ensemble forecast system indicating that,

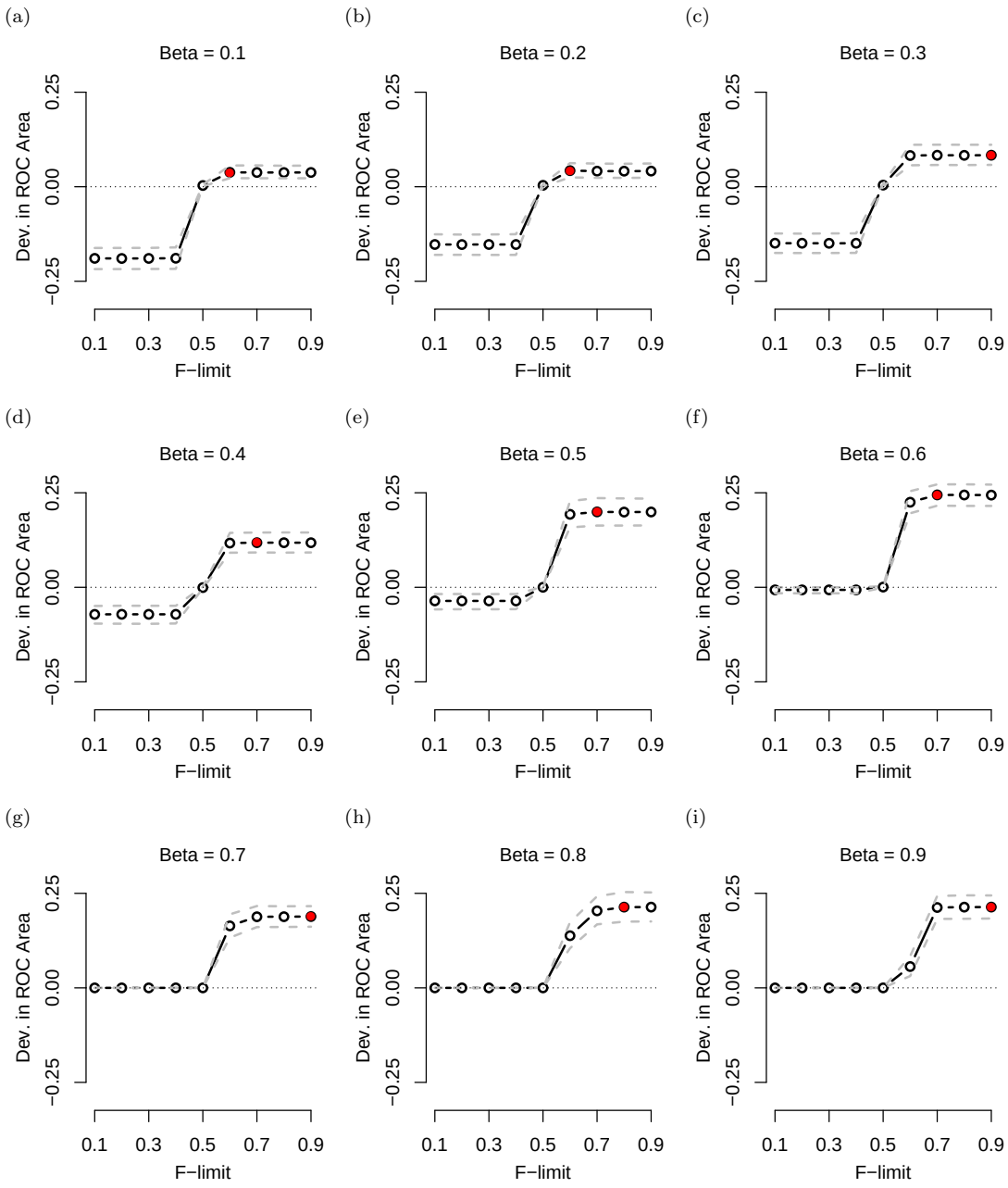

Figure 4.3: Mean deviations from the median f-limit parameter for each value of β. Confidence intervals (gray dashed lines) are empirically derived using a bootstrap algorithm. The red point indicates the maximum positive deviation in ROC area for the given β.

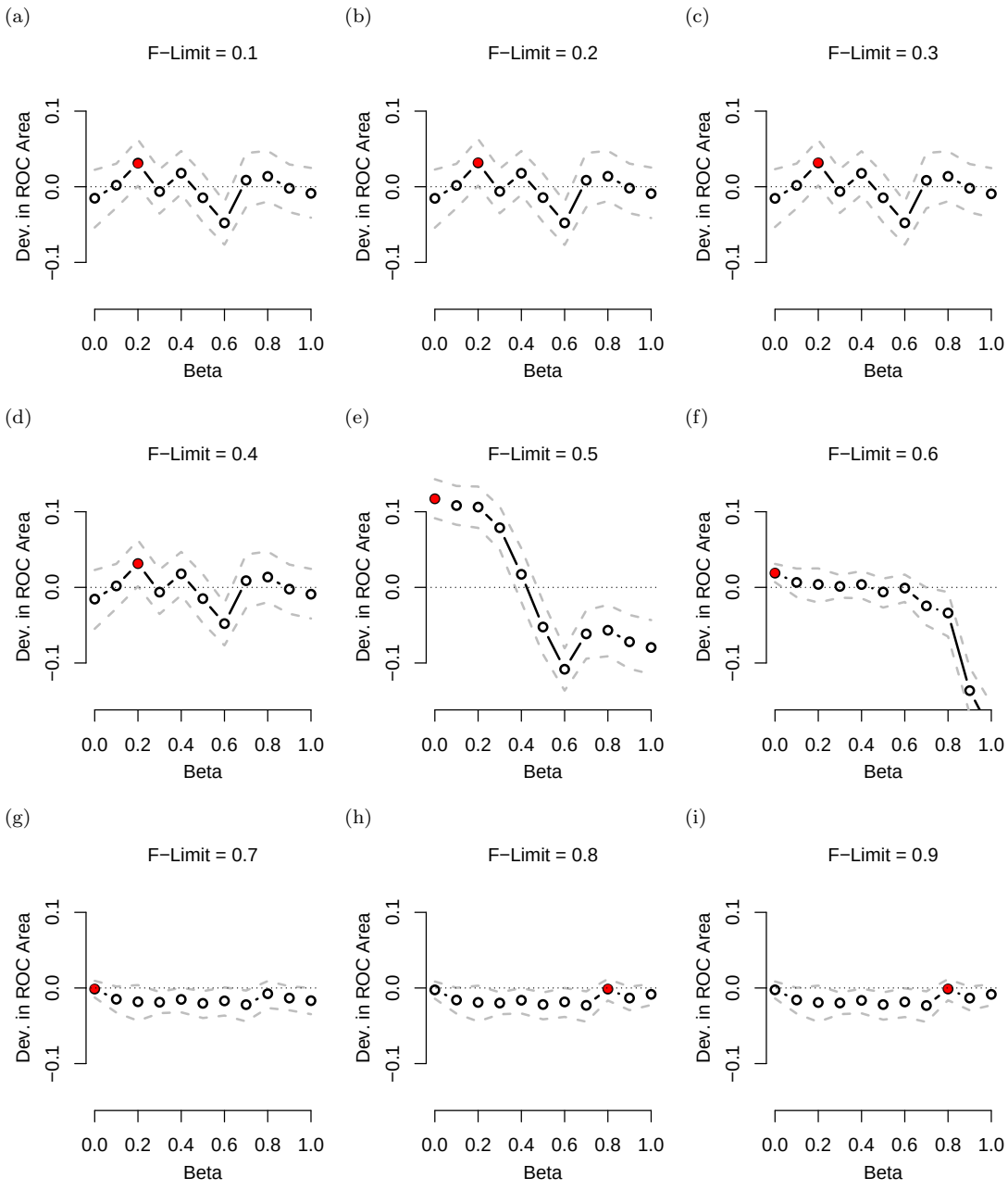

Figure 4.4: Mean deviations from the median β parameter for each value of f-limit. Confidence intervals (gray dashed lines) are empirically derived using a bootstrap algorithm. The red point indicates the maximum positive deviation in ROC area for the given f-limit.

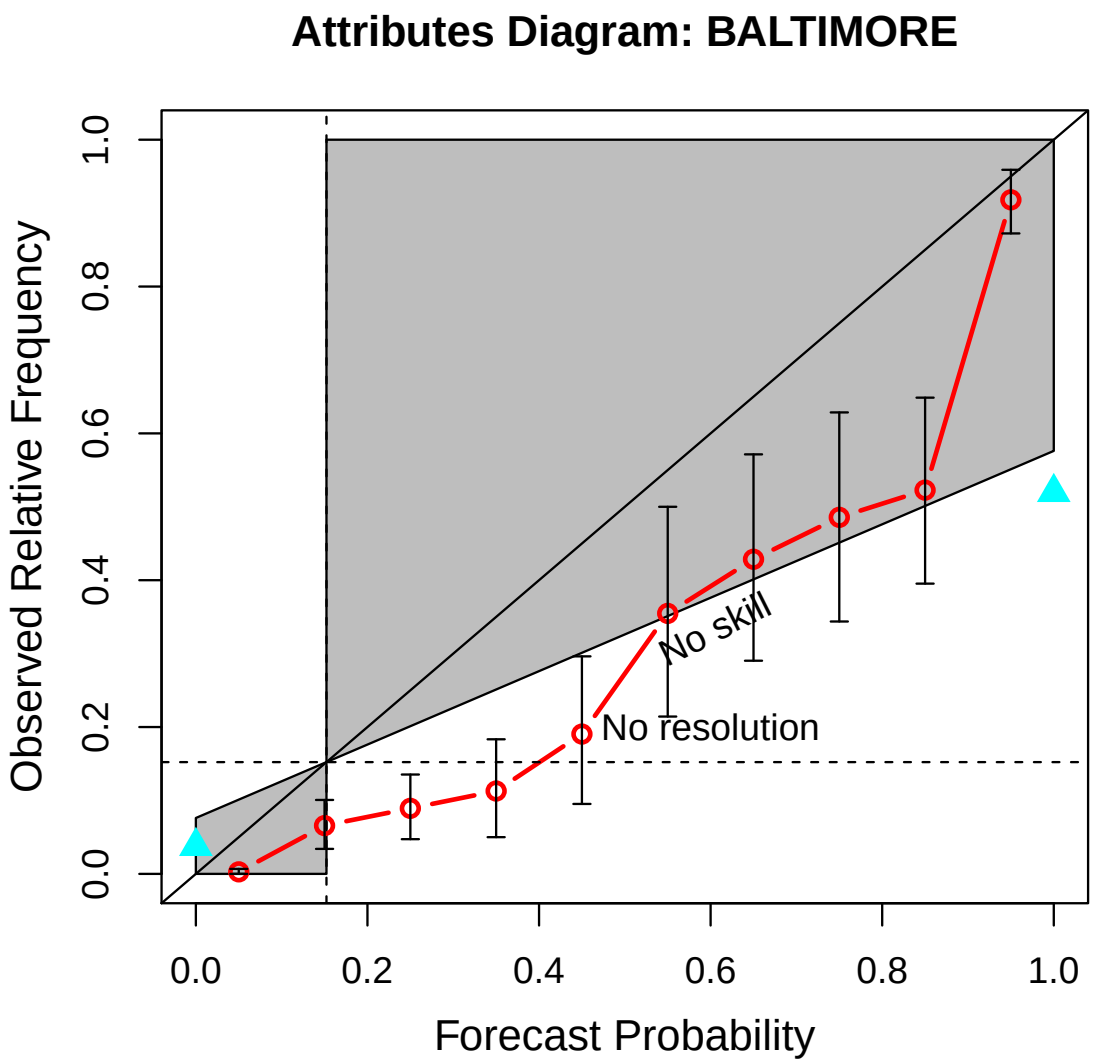

Figure 4.5: Attributes diagram for the ESP product for the Baltimore, MD forecast region. The observed relative frequency of an exceedance event is plotted as a function of the forecasted probability based on the ESP product. An ideal diagram would follow the 1:1 line indicating that the forecasted probability perfectly matches the observed frequency of exceedances given the forecast. The error bars represent the 95% confidence intervals about the mean observed relative frequency for a given forecast probability derived empirically from 10,000 bootstrapped subsamples. The forecast probabilities are binned into 10% bins to provide enough sample points from which to derive confidence intervals as well as facilitate interpretation. The triangles are the NAQFC forecasts converted into binary forecasts for ozone exceedances using the NAAQS threshold of 75 ppbv.

for example, ozone exceeds 30% of the time when the forecasted probability is 30%. A point falling on the "No resolution" line indicates that the associated subset of forecasts are unable to discern events that are different from the climatological probability of the event. The "No skill" line is half-way between the ideal 1:1 line and the "No resolution" line and defines the region where forecasts produce positive skill (shaded) versus negative skill (non-shaded). The points lie below the 1:1 line at 95% significance with the exception of the highest forecast probabilities. This diagram indicates that the ESP tends to overpredict the frequency of ozone exceedances. Observed frequency of ozone exceedances associated with forecast probability between 0.3 and 0.5 are not significantly different from the climatological frequency. Forecast probabilities between 0.5 and 0.9 on average produce marginal positive skill for the ESP, though not statistically significant. Most of the skill of the ESP lies in the lowest (0 - 0.2) and highest (0.9 - 1) forecast probabilities. This conditional bias is expected because the ESP was developed with the intent of forecasting ozone exceedances, thus sacrificing accurate predictions for middle-frequency events. Some of the conditional bias may also be attributed to overfitting during the regression tree training process. The triangles are the NAQFC forecasts converted into binary forecasts of either exceedance (NAQFC $>$ 75 ppbv or probability = 1) or non-exceedance (NAQFC $\leq$ 75 ppbv or probability = 0) and are provided for reference.

The ROC curve (Swets, 1979; Mason, 1982; Wilks, 2011) shown in Fig. 4.6 provides an analysis of forecast performance from the perspective of a decision maker. The ESP hit rate is plotted as a function of the ESP false alarm rate. The probabilistic forecasts from the ESP are converted into binary predictions using a threshold of probability. For example, a threshold of 0.5 means that any forecast with a probability of exceedance greater than or equal to 0.5 would be a forecast for an exceedance while a forecast probability less than 0.5 would be a forecast for a non-exceedance. The hit rate is the proportion of correctly predicted exceedances to the number of observed exceedances while the false alarm rate is the proportion of incorrectly forecasted exceedances to the total number of non-exceedances. Points on the ROC curve are associated with various thresholds of forecast probability that determine a forecast for exceedance. The ideal curve would create a right-angle in the upper-left corner of the plot suggesting there would be a probability threshold that produces a perfect hit rate while never producing a false alarm. Interpretation of this plot depends on the needs of the user. A user whose decision is sensitive to the false alarm rate of the forecast system may use this plot to identify a forecast probability threshold that maximizes the hit rate while remaining below an acceptable false alarm rate. For example, such a user would be able to achieve a 92% hit rate while remaining below the 20% false alarm rate using the ESP and a probability threshold of 35%. Compared to the NAQFC, the ESP produces a 7.5% increase in hit rate at the same false alarm rate and a 4.2% decrease in false alarm rate at the same hit rate. The distribution of ROC area (area under the ROC curve) for the ESP is tightly centered around a mean of 0.95 and is significantly different from the ROC area associated with the NAQFC (vertical dashed line).

The goal of developing the ESP was to address the difficulties in forecasting ozone in Baltimore while providing increased value of information. Fig. 4.7 depicts the increase in value of the air

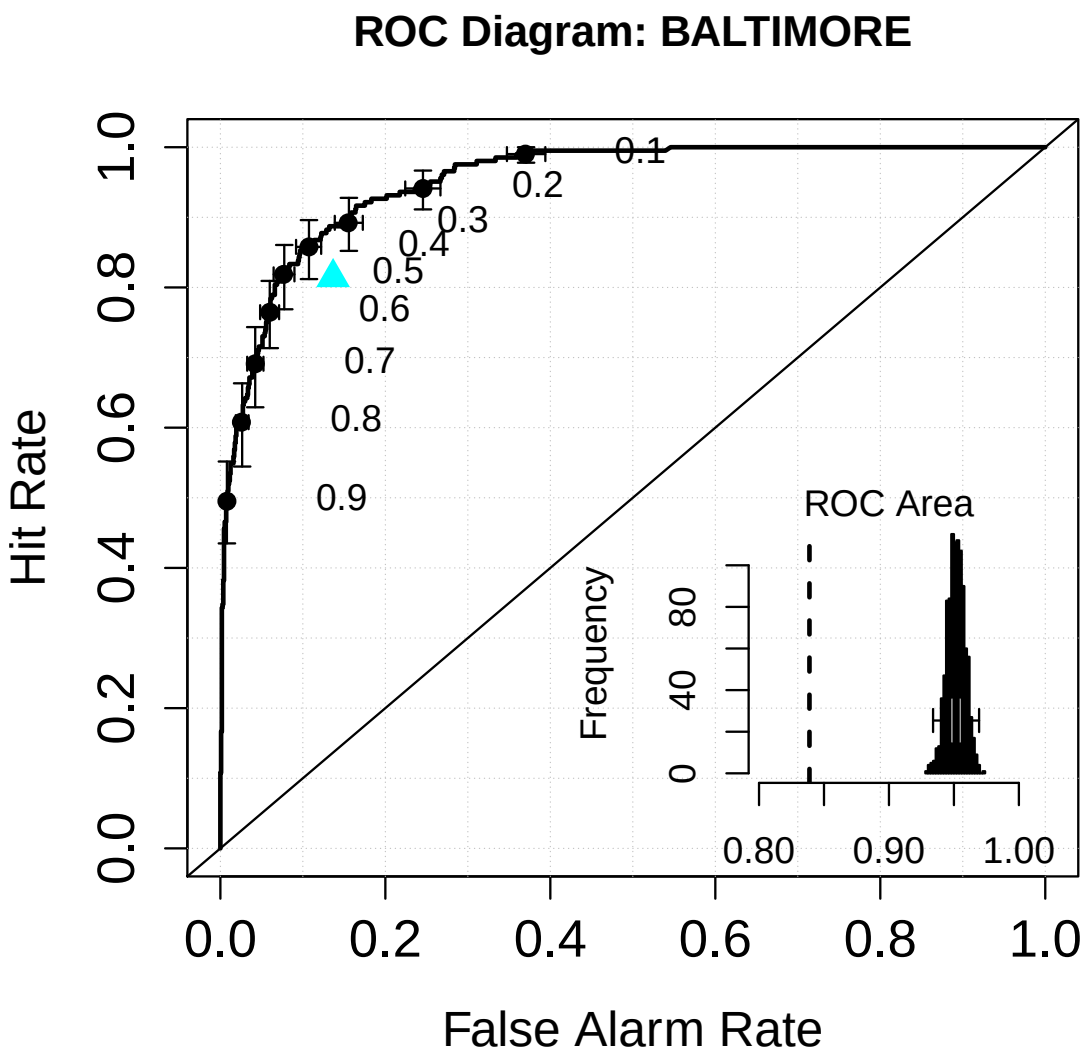

Figure 4.6: Relative operating characteristic (ROC) diagram for the ESP product for the Baltimore, MD forecast region. The hit rate is plotted as a function of the false alarm rate for a series of forecast probability thresholds which define a forecasted exceedance. The dots represent the different forecast probability thresholds used to convert the probabilistic forecast into a binary forecast. The error bars are the 95% confidence interval about the mean hit rate (vertical) and false alarm rate (horizontal) derived empirically from 10,000 bootstrap subsamples. The triangle is the NAQFC forecast. The histogram inset describes the distribution of the area under the ROC curve based on the bootstrap subsamples used in deriving the confidence intervals. The vertical dashed line in the inset represents the area of the ROC curve based on the NAQFC forecast.

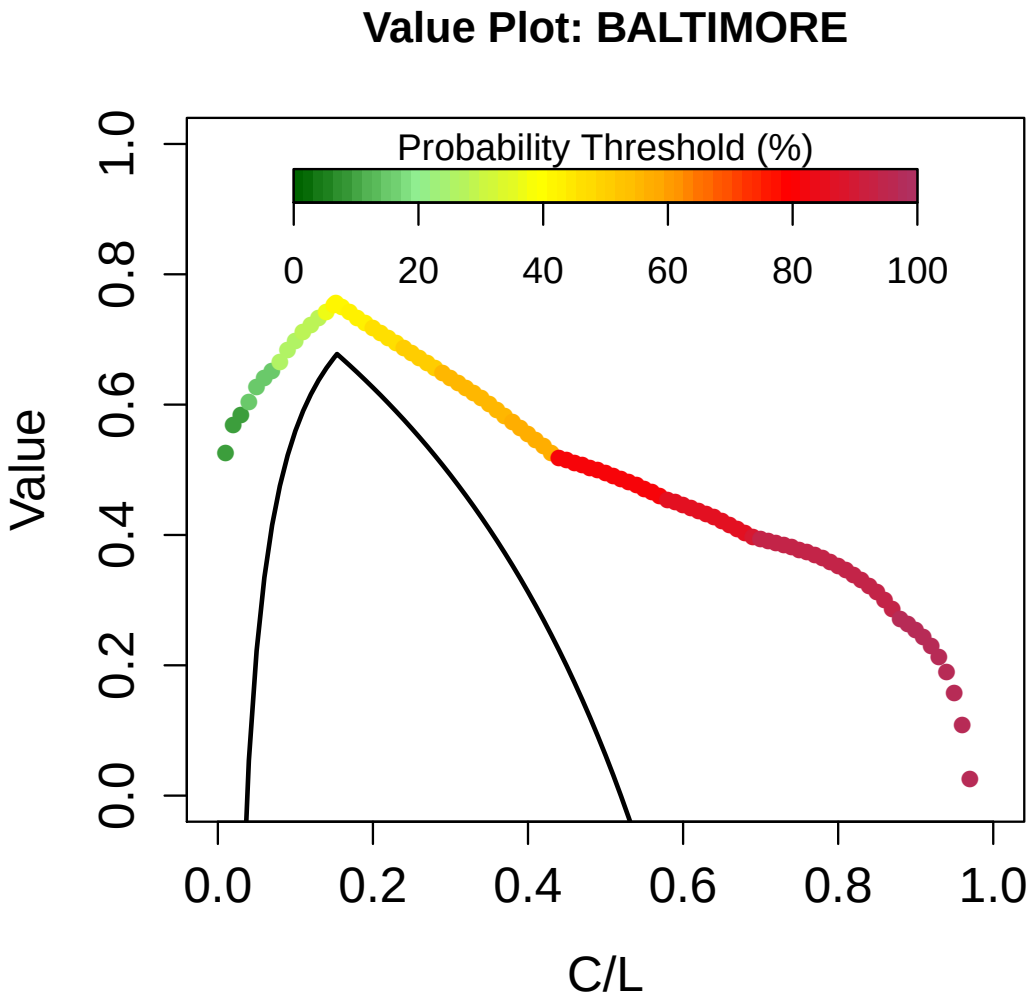

Figure 4.7: Value curve for the ESP product for the Baltimore, MD forecast region. Color shading represents the probability threshold used to get the maximum value for the given cost-loss ratio. The solid black line is the value curve for the NAQFC.

quality forecasts when using the ESP relative to the NAQFC alone. Value is calculated using a static cost-loss ratio model (Thompson, 1952; Richardson, 2000; Garner and Thompson, 2012; Garner et al., 2013). Value is the percent savings in expenditure over climatological expenditure relative to a perfect forecast system. The cost-loss ratio (C/L) is defined by the cost (C) to insure against loss (L) of an ozone exceedance. For a given loss, C/L can be interpreted as a suite of possible decisions relevant to the user. For example, the maximum value of the ESP is 0.75 at a C/L of 0.18 using a probability threshold of 44%. This means that if the user makes a decision based on the ESP to implement a program that costs \$18,000 to prevent a loss of \$100,000, the

user can expect to save \$11,070 per event [$18\% \times (\$100,000 - \$18,000) \times 75\% = \$11,070$]. The benefit of an ensemble forecast system is that the user can maximize the value of the forecasts for multiple decisions by choosing multiple probability thresholds that define a forecasted exceedance. Say that the same user from before would like to attain the maximum value for their decision costing \$60,000. This user would use a probability threshold of 84% for this decision as opposed to the 44% threshold used earlier. The ESP not only provides an increase in value of information over the NAQFC alone (black curve in Fig. 4.7), but the ensemble approach also provides value at C/L far exceeding what the NAQFC could provide. This means that the NAQFC, after application of the ESP, can not only provide more value for decisions currently covered by the NAQFC, but also value for high-cost decisions currently not covered by the NAQFC.

4.4 Summary and Discussion

The ESP was developed to address the challenges in forecasting air quality in Baltimore, MD. These challenges include local meteorological phenomena unresolvable by the current numerical models, uncertainty about air quality forecasts, and pitfalls in statistical assumptions when fitting standard statistical models.

Ozone and meteorological data were collected from eight ozone monitors that represent the Baltimore forecast region. These data were used to develop 100 regression trees for each monitor using a moving-block bootstrap algorithm. The 100 regression trees constitute the ESP for the given monitor. Each regression tree is fitted using the f-measure (Eq. 4.1) for node splitting and evaluation. The result is regression trees with meteorology-dependent paths leading to nodes tailored to predicting ozone exceedances. The ESP was evaluated with a 10-fold cross-validation designed to mimic an operational forecasting scenario.

Results indicate that the ESP exhibits conditional bias. This conditional bias was expected due to the measures taken to achieve the goal of the product. In addition, some overfitting of the regression tree models may have contributed to the conditional bias. The ESP tends to overpredict the relative frequency of ozone exceedances between forecast probabilities of 0.2 - 0.9. The skill of the ESP is shown at the extremes for forecast probabilities. Individual monitors behave better in this regard when compared to the region as a whole.

From the perspective of a decision maker, the ESP provides significantly more information than the NAQFC alone. The ESP allows the user to choose probability thresholds that fit the criteria of their decision, such as sensitivity to false alarms or minimum achievable hit rate. The ESP was also shown to not only improve the value of the NAQFC, but also provide significant value at decisions not attainable with the NAQFC alone.

The ESP can be a valuable tool to an air quality forecaster. This article describes the ESP development for forecasting ozone in Baltimore, but these methods can be easily adapted and applied to many other forecast challenges. The authors have begun and would like to continue with the operational implementation of the ESP in order to assess its performance in true operational scenarios.

Chapter 5

Summary

The goal of this research was to develop a tool that aids air quality forecasting and decision-making regarding ozone in the mid-Atlantic region. To accomplish this goal, this research was divided into three tasks.

The first task was to establish a baseline by analyzing the value of information that current forecast systems provide. Forecasts from three different forecast systems were analyzed for skill and value of information. Forecast skill was measured using standard continuous and categorical statistics while value of information was measured using a static cost-loss ratio model. Results indicated that human forecasters provide significant value of information across a broad range of potential decisions, however, less accurate forecast systems can provide more valuable information than human forecasters for specific decisions.

The second task was to gain a sense of the current trend in air quality forecast value by assessing the next-generation of operational numerical air quality models. Forecasts from the NAQFC were compared to the NAQFC-β, an experimental version of the NAQFC with an upgraded chemical mechanism, for both skill and value during a NASA field campaign in July 2011. Statistically significant biases were found in both the NAQFC and the NAQFC-β using empirical distributions of error derived from bootstrap subsampling. While the NAQFC tended to outperform the NAQFC-β, the NAQFC-β provided up to 70% more value for low cost-loss ratio decisions.

The final task was to develop the ESP so that it improved upon the value of information already provided by the NAQFC. Ozone and meteorological data from the eight monitors defining the Baltimore forecast region during the ozone seasons of 2005 - 2011 were used to develop this tool. The ESP development involved training a series of regression trees based on bootstrap subsamples of the original data. This allowed the final product to capture the uncertainties arising during the regression tree training process. The f-measure was used as the evaluation function within the regression trees, tailoring each tree to predicting ozone exceedance events (defined by the NAAQS for ozone of 75 ppbv for an 8-h average). The result was eight site-specific ESPs, each a set of 100 regression tree models designed to capture the meteorological conditions conducive

to ozone exceedance events. The ESP was evaluated through a 10-fold cross-validation process which mimics evaluation under operational forecasting conditions. Evaluation results indicate that the ESP is slightly conditionally biased, likely due to some overfitting during the development process. The ESP does, however, provide significant improvements in the value and utility of the NAQFC forecasts. Not only does the ESP increase the value of NAQFC forecasts in the Baltimore region, but it also provides value for decisions which the NAQFC alone could not.

Operational testing of the ESP has begun using the Short-Range Ensemble Forecast (SREF) system as input into the ESP. The ESP output is shared through a web-based interface with regional air quality forecasters. The air quality forecasters responded well to the ESP, appreciating its ability to convey the uncertainties in air quality forecasting. Through the feedback from the air quality forecasters, the ESP was expanded to include over 40 different monitors throughout the mid-Atlantic.

The ESP must be run for about two to three complete ozone seasons in order to compile a data set from which metrics regarding the skill and value of the ESP in true operational forecasting can be calculated. To continue this work, one would need to run the ESP for the next few ozone seasons and perform the same analyses described in this work. Additionally, an ESP can be trained for particle pollution forecasts based on the NAQFC-β for regions where particle pollution is the main air quality issue. Such an application of ESP could yield increases in value that potentially surpass the increase in value of the ESP for ozone forecasts.

Bibliography

Al-Alawi, S.M., Abdul-Wahab, S.A., Bakheit, C.S., 2008. Combining principal component regression and artificial neural networks for more accurate predictions of ground-level ozone. Environmental Modelling and Software 23, 396 – 403, DOI:10.1016/j.envsoft.2006.08.007.

Anderson, R.C., 2001. Pollution charges, fees, and taxes, in: The U.S. Experience with Economic Incentives for Protecting the Environment. Environmental Protection Agency: National Center for Environmental Economics, pp. 33–55.

Aron, R., Aron, I., 1978. Statistical forecasting models: Carbon monoxide concentrations in the Los Angeles Basin. J. Air Pollut. Control Assoc. 28, 681–684.

Banta, R.M., Senff, C.J., Neilson-Gammon, J., Darby, L.S., Ryerson, T.B., Alvarez, R.J., Sandberg, S.P., Williams, E.J., Trainer, M., 2005. A bad air day in houston. Bull. Amer. Meteor. Soc. 86, 657 – 69, DOI:10.1175/BAMS-86-5-657.

Berger, J.O., 2006. Basic concepts, in: Statistical Decision Theory and Bayesian Analysis. Springer Science+Business Media LLC, 2 edition. pp. 1–34.

Berry, M., Lioy, P.J., Gelperin, K., Buckler, G., Klotz, J., 1991. Accumulated exposure to ozone and measurement of health effects in children and counselors at two summer camps. Environ. Research 54, 135–150.

Bey, I., Jacob, D.J., Yantosca, R.M., Logan, J.A., Field, B.D., Fiore, A.M., Li, Q., Liu, H.Y., Mickley, L.J., Schultz, M.G., 2001. Global modeling of tropospheric chemistry with assimilated meteorology: Model description and evaluation. Journal of Geophysical Research 106, 23073 – 95.

Breiman, L., 1984. Classification and regression trees. Wadsworth statistics/probability series, Wadsworth International Group.

Byun, D., Schere, K.L., 2006. Review of the governing equations, computational algorithms, and other components of the models-3 Community Multiscale Air Quality (CMAQ) modeling system. Applied Mechanics Reviews 59, 51–77.

Chen, T., Gokhale, J., Shofer, S., Kuschner, W.G., 2007. Outdoor air pollution: Ozone health effects. The American J. of the Medical Sciences 333, 244–248.

Davis, J.M., Speckman, P., 1999. A model for predicting maximum and 8 h average ozone in houston. Atmos. Environ. 33, 2487–2500.

Dimitriades, B., 1976. Photochemical Oxidants in the Ambient Air. Technical Report EPA-600/3-76-017. Environmental Protection Agency.

Djalalova, I., Wilczak, J., McKeen, S., Grell, G., Peckham, S., Pagowski, M., DelleMonache, L., McQueen, J., Tang, Y., Lee, P., McHenry, J., Gong, W., Bouchet, V., Mathur, R., 2010. Ensemble and bias-correction techniques for air quality model forecasts of surface o3 and pm2.5 during the texaqs-ii experiment of 2006. Atmospheric Environment 44, 455 – 467, DOI:10.1016/j.atmosenv.2009.11.007.

Eder, B., Kang, D., Mathur, R., Pleim, J., Yu, S., Otte, T., Pouliot, G., 2009. A performance evaluation of the national air quality forecast capability for the summer of 2007. Atmos. Environ. 43, 2312–2320.

Eder, B., Kang, D., Mathur, R., Yu, S., Schere, K., 2006. An operational evaluation of the Eta-CMAQ air quality forecast model. Atmos. Environ. 40, 4894–4905.

Eder, B., Kang, D., Rao, S.T., Mathur, R., Yu, S., Otte, T., Schere, K., Wayland, R., Jackson, S., Davidson, P., McQueen, J., Bridgers, G., 2010. Using national air quality forecast guidance to develop local air quality index forecasts. Bulletin of the American Meteorological Society 91, 313–326, DOI:10.1175/2009BAMS2734.1.

Efron, B., 1979. Bootstrap methods: Another look at the jackknife. The Annals of Statistics 7, pp. 1–26.

Efron, B., Gong, G., 1983. A leisurely look at the bootstrap, the jackknife, and cross-validation. The American Statistician 37, pp. 36–48.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman and Hall.

Foley, K.M., Roselle, S.J., Appel, K.W., Bhave, P.V., Pleim, J.E., Otte, T.L., Mathur, R., Sarwar, G., Young, J.O., Gilliam, R.C., Nolte, C.G., Kelly, J.T., Gilliland, A.B., Bash, J.O., 2010. Incremental testing of the community multiscale air quality (CMAQ) modeling system version 4.7. Geoscientific Model Development 3, 205–226, DOI:10.5194/gmd-3-205-2010.

Garner, G.G., Thompson, A.M., 2012. The value of air quality forecasting in the mid-atlantic region. Wea. Climate Soc. 4, 69–79, DOI:10.1175/WCAS-D-10-05010.1.

Garner, G.G., Thompson, A.M., Lee, P., Martins, D.K., 2013. Evaluation of naqfc model performance in forecasting surface ozone during the 2011 discover-aq campaign. Journal of Atmospheric Chemistry , 1–19, DOI:10.1007/s10874-013-9251-z.

Glahn, H.R., Lowry, D.A., 1972. The use of model output statistics *mos* in objective weather forecasting. J. Appl. Meteor. 11, 1203–11.

Houyoux, M.R., Vukovich, J.M., Coats Jr., C.J., Wheeler, N.J.M., Kasibhatla, P.S., 2000. Emission inventory development and processing for the seasonal model for regional air quality (SMRAQ) project. J. Geophys. Res. 105, 9079–9090.

Hsu, W.R., Murphy, A.H., 1986. The attributes diagram a geometrical framework for assessing the quality of probability forecasts. International Journal of Forecasting 2, 285 – 293, DOI:10.1016/0169-2070(86)90048-8.

Hubbard, M.C., Cobourn, W.G., 1998. Development of a regression model to forecast ground-level ozone concentrations in Jefferson County, Kentucky. Atmos. Environ. 32, 2637–2647.

Janjic, Z.I., 2003. A nonhydrostatic model based on a new approach. Meteor. Atmos. Phys. 82, 271–285.

Johnson, D.L., Ambrose, S.H., Bassett, T.J., Bowen, M.L., Crummey, D.E., Isaacson, J.S., Johnson, D.N., Lamb, P., Saul, M., Winter-Nelson, A.E., 1997. Meanings of environmental terms. J. Environ. Quality 26, 581–89.

Katz, R.W., Murphy, A.H., 1997. Forecast value: prototype decision-making models, in: Economic Value of Weather and Climate Forecasts. Cambridge University Press, pp. 183–217.

Kernan, G.L., 1975. The cost-loss decision model and air pollution forecasting. J. Appl. Meteor. 14, 8–16.

Klein, W.H., Lewis, B.M., Enger, I., 1959. Objective prediction of five-day mean temperatures during winter. Journal of Meteorology 16, 672 – 82.

Krupnick, A.J., Ostro, W.H.B., 1990. Ambient ozone and acute health effects: Evidence from daily data. J. Environ. Econom. Manage. 18, 1–18.

Lee, P., Ngan, F., 2011. Coupling of important physical processes in the planetary boundary layer between meteorological and chemistry models for regional to continental scale air quality forecasting: An overview. Atmosphere 2, 464–483, DOI:10.3390/atmos2030464.

Lin, Y., 1982. Oxidant prediction by discriminant analysis in the south coast air basin of California. Atmos. Environ. 16, 135–143.

Lippmann, M., 1989. Health effects of ozone. a critical review. J. Air. Pollut. Control Waste Manage. 39, 672–695.

Loughner, C.P., Allen, D.J., Pickering, K.E., Zhang, D.L., Shou, Y.X., Dickerson, R.R., 2011. Impact of fair-weather cumulus clouds and the chesapeake bay breeze on pollutant transport and transformation. Atmospheric Environment 45, 4060 – 4072, DOI:10.1016/j.atmosenv.2011.04.003.

Mason, I., 1982. A model for assessment of weather forecasts. Aust. Met. Mag 30, 291–303.

McCollister, G.M., Wilson, K.R., 1975. Linear stochastic models for forecasting daily maxima and hourly concentrations of air pollutants. Atmos. Environ. 9, 417–423.

Millner, A., 2009. What is the true value of forecasts? Weather, Climate, and Society 1, 22–37.

Mintz, D., 2009. Technical Assistance Document for the Reporting of Daily Air Quality: The Air Quality Index (AQI). Technical Report EPA-454/B-09-001. Environmental Protection Agency.

Nolen, J.E., Billings, P.G., Sukachevin, N., Iwanowicz, P., Moseley, L., Jump, Z., Lancet, E., Rappaport, S., Edelman, N., 2013. State of the Air 2013. Technical Report. American Lung Association.

Picard, R.R., Cook, R.D., 1984. Cross-validation of regression models. Journal of the American Statistical Association 79, pp. 575–583.

Pires, J., Martins, F., 2011. Correction methods for statistical models in tropospheric ozone forecasting. Atmospheric Environment 45, 2413 – 2417, DOI:10.1016/j.atmosenv.2011.02.011.

Ribeiro, R., Torgo, L., 2006. Rule-based prediction of rare extreme values, in: Todorovski, L., Lavrac, N., Jantke, K. (Eds.), Discovery Science. Springer Berlin / Heidelberg. volume 4265 of *Lecture Notes in Computer Science*, pp. 219–230.

Ribeiro, Jr., P.J., Diggle, P.J., 2001. geoR: A package for geostatistical analysis. R News 1, 14–18.

Richardson, D.S., 2000. Skill and relative economic value of the ECMWF ensemble prediction system. Q. J. R. Meteorol. Soc. 126, 649–67.

Robeson, S.M., Steyn, D.G., 1990. Evaluation and comparison of statistical forecast models for daily maximum ozone concentrations. Atmos. Environ. 24B, 303–312.

Rotach, M.W., Ambrosetti, P., Appenzeller, C., Arpagaus, M., Fontannaz, L., Fundel, F., Germann, U., Hering, A., Liniger, M.A., Stoll, M., Walser, A., Ament, F., Bauer, H.S., Behrendt, A., Wulfmeyer, V., Bouttier, F., Seity, Y., Buzzi, A., Davolio, S., Corazza, M., Denhard, M., Dorninger, M., Gorgas, T., Frick, J., Hegg, C., Zappa, M., Keil, C., Volkert, H., Marsigli, C., Montaini, A., McTaggart-Cowan, R., Mylne, K., Ranzi, R., Richard, E., Rossa, A., Santos-Muoz, D., Schr, C., Staudinger, M., Wang, Y., Werhahn, J., 2009. Map D-PHASE: Real-time demonstration of weather forecast quality in the alpine region. Bulletin of the American Meteorological Society 90, 1321–1336, DOI:10.1175/2009BAMS2776.1.

Ryan, W.F., 1995. Forecasting severe ozone episodes in the Baltimore metropolitan area. Atmos. Environ. 29, 2387–2398.

Ryan, W.F., Davidson, P., Stokols, P., Carey, K., 2004. Evaluation of the National Air Quality Forecast System (NAQFS): Summary of the air quality forecasters focus group workshop, in: Air Quality in Megacities (Joint Symposium on Planning, Nowcasting and Forecasting in the Urban Zone and 6^{th} Conference on Atmospheric Chemistry, American Meteorological Society, Seattle, WA. p. J2.13.

Sarwar, G., Luecken, D., Yarwood, G., Whitten, G.Z., Carter, W.P.L., 2008. Impact of an updated carbon bond mechanism on predictions from the cmaq modeling system: preliminary assessment. Journal of Applied Meteorology and Climatology 47, 3–14.

Saylor, R.D., Stein, A.F., 2012. Identifying the causes of differences in ozone production from the CB05 and CBMIV chemical mechanisms. Geoscientific Model Development 5, 257–268, DOI:10.5194/gmd-5-257-2012.

Seinfeld, J.H., Pandis, S.N., 2006. Chemistry of the Troposphere, in: Atmospheric Chemistry and Physics: From Air Pollution to Climate Change. John Wiley and Sons Inc., 2 edition. pp. 205–283.

Sikora, T.D., Young, G.S., Bettwy, M.J., 2010. Analysis of the western shore chesapeake bay bay-breeze. National Weather Digest 34, 55 – 65.

Stauffer, R.M., Thompson, A.M., 2013. Bay breeze climatology at two sites along the chesapeake bay from 1986-2010: Implications for surface ozone. Journal of Atmospheric Chemistry , 1–18, DOI:10.1007/s10874-013-9260-y.

Stauffer, R.M., Thompson, A.M., Martins, D.K., Clark, R.D., Goldberg, D.L., Loughner, C.P., Delgado, R., Dickerson, R.R., Stehr, J.W., Tzortziou, M.A., 2012. Bay breeze influence on surface ozone at edgewood, md during july 2011. Journal of Atmospheric Chemistry , 1–19, DOI:10.1007/s10874-012-9241-6.

Swets, J.A., 1979. ROC analysis applied to the evaluation of medical imaging techniques. Investigative radiology 14, 109–121.

Tang, Y., Lee, P., Tsidulko, M., Huang, H., McQueen, J.T., DiMego, G.J., Emmons, L.K., Pierce, R.B., Thompson, A.M., Lin, H., Kang, D., Tong, D., Yu, S., Mathur, R., Pleim, J.E., Otte, T.L., Pouliot, G., Young, J.O., Schere, K.L., Davidson, P.M., Stajner, I., 2008. The impact of chemical lateral boundary conditions on cmaq predictions of tropospheric ozone over the continental united states. Environ. Fluid Mech. , DOI:10.1007/s10652-008-9092-5.

Thompson, J.C., 1952. On the operational deficiencies in categorical weather forecasts. Bull. Amer. Meteor. Soc. 33, 223–226.

Thompson, J.C., Brier, G.W., 1955. The economic utility of weather forecasts. Mon. Wea. Rev. 83, 249–254.

Thompson, M.L., Reynolds, J., Cox, L.H., Guttorp, P., Sampson, P.D., 2001. A review of statistical methods for the meteorological adjustment of tropospheric ozone. Atmospheric Environment 35, 617 – 630, DOI:10.1016/S1352-2310(00)00261-2.

Thornes, J.E., 2001. How to judge the quality and value of weather forecast products. Journal of Applied Meteorology 8, 307–314.

Torgo, L., Ribeiro, R., 2003. Predicting outliers, in: Lavrac, N., Gamberger, D., Todorovski, L., Blockeel, H. (Eds.), Knowledge Discovery in Databases: PKDD 2003. Springer Berlin / Heidelberg. volume 2838 of *Lecture Notes in Computer Science*, pp. 447–458.

Wandishin, M.S., Brooks, H.E., 2002. On the relationship between clayton's skill score and expected value for forecasts of binary events. Meteorol. Appl. 9, 455–9, DOI:10.1017/S1350482702004085.

Wilks, D.S., 2011. Statistical Methods in the Atmospheric Sciences. Elsevier. 3 edition.

Wright, E.S., Dziedzic, D., Wheeler, C.S., 1990. Cellular, biochemical and functional effects of ozone: new research and perspectives on ozone health effects. Toxicology Letters 51, 125–145.

Yarwood, G., Rao, S., Yocke, M., Whitten, G.Z., 2005. UPDATES TO THE CARBON BOND CHEMICAL MECHANISM: CB05. Technical Report RT-04-00675. Yocke and Company.

Vita

Gregory George Garner

Gregory George Garner was born March 13, 1984 in Boston, MA. He attended the Boston Latin School, where he focused his education on the classics (Latin and Declamation), science (Environmental Science and Physics), and technology (Computer Programming). This was when his passion for Meteorology was born. In 2002, he graduated from the Boston Latin School and began his study of Meteorology at what was then Plymouth State College.

Greg was fortunate to participate in a summer internship at the university to study New England air quality. He took a leadership role and organized the investigation of eight case-studies with his fellow interns. He felt that air quality will be of great interest in light of concerns regarding climate change and so air quality became the primary focus of his work. In 2006, he graduated with honors from the Meteorology program at what has become Plymouth State University with minors in Technical Mathematics and Physics.

Greg continued to pursue a masters degree in the newly founded Applied Meteorology program at Plymouth State University. During this time, he tutored undergraduates in Meteorology and Physics while maintaining a job at a local grocery store. He defended a thesis on methods of forecasting air quality in New Hampshire and graduated in 2007. The following spring semester, he took an adjunct faculty position at Plymouth State University, teaching "Introduction to Weather" while preparing to attend The Pennsylvania State University in the fall.

Greg joined an air quality research group at Penn State in 2008 in order to pursue a doctoral degree in Meteorology. He was awarded the Anne C. Wilson Graduate Reseach Award in 2008 and an EPA Science to Achieve Results (STAR) fellowship in 2011. He has presented his work at numerous regional and national conferences, including American Meteorological Society annual meetings and the American Geophysical Union Conference. He worked closely with Maryland and Virginia air quality forecasters in order to develop new forecast tools that improve the value of air quality forecasts in the mid-Atlatic. This work became the focus and drive of his PhD dissertation.

Greg is also a member of the American Meteorological Society (2003), US Mensa (2006), and the American Geophysical Union (2012). His hobbies include building and coding computer projects, basketball, and photography.