The Pennsylvania State University

The Graduate School

College of Earth and Mineral Sciences

PROBABILISTIC FORECASTING OF SURFACE OZONE WITH A NOVEL STATISTICAL APPROACH

A Dissertation in

Meteorology

by

Nikolay Balashov

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2016

The dissertation of Nikolay Balashov was reviewed and approved* by the following:

Anne M. Thompson
Adjunct Professor of Meteorology
Dissertation Advisor
Co-Chair of Committee

George Young
Professor of Meteorology and Geo-Environmental Engineering
Co-Chair of Committee

Steven Greybush
Assistant Professor of Meteorology

Benjamin Shaby
Assistant Professor of Statistics

William Ryan
Adjunct Research Assistant
Special Member

Johannes Verlinde
Professor of Meteorology
Associate Head, Graduate Program in Meteorology

*Signatures are on file in the Graduate School

ABSTRACT

The recent change in US Environmental Protection Agency (EPA) surface ozone regulation, lowering surface ozone daily maximum 8-hour average (MDA8) exceedance threshold from 75 ppbv to 70 ppbv, poses significant challenges to US air quality (AQ) forecasters responsible for ozone MDA8 forecasts. The forecasters, supplied by only a few AQ model products, end up relying heavily on self-developed tools. To help US AQ forecasters, this study explores surface ozone MDA8 forecasting tool based solely on statistical methods and standard meteorological variables from the numerical weather prediction (NWP) models. The model combines self-organizing map (SOM), a clustering technique, with a stepwise weighted quadratic regression using meteorological variables as predictors for ozone MDA8. The SOM method identifies different weather regimes, to distinguish between various modes of ozone variability, and groups them according to similarity. In this way, when a regression is developed for a specific regime, data from the other regimes are also used, with weights based on their similarity to this specific regime. This approach, regression in SOM (REGiS), yields a distinct model for each regime taking into account both the training cases for that regime and other similar training cases. To produce probabilistic MDA8 ozone forecasts, REGiS weighs and combines all of the developed regression models based on the weather patterns predicted by a NWP model. REGiS is evaluated over San Joaquin Valley in California and northeastern plains of Colorado. The results suggest that the model performs best when trained and adjusted separately for an individual AQ station and its corresponding meteorological site. Real-time ozone forecasting using REGiS is demonstrated for the Philadelphia area over a brief period of time in 2016.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This work would not have been possible without the constant support and encouragement from my parents. Of course I would like to acknowledge my patient committee. Important acknowledgments go to 7th International Workshop on Air Quality Forecasting Research (IWAQFR), held in College Park, MD, in 2015, for the significant input regarding the state of international air quality forecasting in 2015. Special thanks to air quality operational forecasters William Ryan, Amy Huff and Joel Dreessen for their feedback regarding the air quality forecasting work flow. Thanks to NASA for their enlightening DISCOVER-AQ campaign experience. Many thanks to Ryan Stauffer, Hannah Halliday, Greg Garner and the rest of the Gator research group for their continued help and support. Finally, much gratitude to all of my friends who supported me in the process of writing this dissertation. This research was supported by NASA through Grants to Penn State University, DISCOVER-AQ (NNX10AR39G) and Applied Sciences Air Quality (NNX11AQ44G).

Chapter 1

Introduction

A string of deadly air pollution episodes during the 1930s-1960s in North America and Europe prompted a creation of air pollution regulations, which in turn led to the development of a research field known today as the air quality (AQ) forecasting, established to protect public (Zhang et al. 2012a). And although the AQ forecasting alone cannot cope with the air pollution problem, it can, in many cases, lead to better decision making and reduce losses (Garner and Thompson 2013). An array of forecasting methods to predict AQ was developed, ranging from statistical models to advanced, physically-based approaches. Despite the efforts of scientists from multiple disciplines such as meteorology, atmospheric chemistry, mathematics, statistics, computer sciences, etc., the new discipline of AQ forecasting met exceptional challenges (Zhang et al. 2012a). The ever-changing chemical composition of the atmosphere requires constant and careful updates to the AQ models, which in turn require proper expertise (Comrie 1997; Zhang et al. 2012a). Given the scope of the air pollution issue, diverse AQ model development is an important task that can help with keeping society well informed about the real-time global air pollution.

Although the AQ has improved significantly in North America and Europe over the past 30 years, the same cannot be said about many other places around the globe, especially developing countries where air pollution is on the rise (Shahgedanova and Burt 1994; Fenger 2009; Klimont et al. 2009; Lamarque et al. 2010). Regions with improved or worsened AQ alike are faced with the issue of constantly changing emissions (Zhang et al. 2012a). That means that the AQ models also need to change continuously. Therefore, creating a model that can quickly adapt to environmental variability is a desirable objective in the field of AQ forecasting.

In this thesis, a computationally fast and flexible model for surface ozone prediction is presented based on a combination of already well established statistical forecasting approaches (Thompson et al. 2001; Zhang et al. 2012a).

1.1 Ozone Pollution

A number of studies have shown damaging effects of air pollution on humans, wildlife, and materials (Richkind and Hacker 1980; Lee et al. 1996; Greenbaum et al. 2001; Brook et al. 2002). In the United States (US), the Environmental Protection Agency (EPA) has established six criteria pollutants that are constantly monitored to protect public health. Of these pollutants, the ones that most commonly exceed unhealthy thresholds are surface ozone and particulate matter (PM) with aerodynamic diameters less than or equal to 2.5 µm ($PM_{2.5}$) and 10 µm (PM_{10}) (EPA 2003). The focus of this thesis, however, is the prediction of the surface ozone.

1.1.1 Ozone Genesis

Ozone is not emitted into the atmosphere of the Earth, but is produced from various precursors by several different chemical mechanisms. Stratospheric ozone forms through the dissociation of molecular oxygen by solar ultraviolet (UV) radiation and subsequent reaction of oxygen atoms with O_2 in the presence of another molecule M (commonly N_2 or O_2). In the troposphere, ozone is accumulated by the downward transport form the stratosphere and is generated through the complex chain of reactions mainly between nitrogen oxides (NO_x), carbon monoxide (CO), methane (CH_4), and volatile organic compounds (VOCs) in the presence of sunlight. NO_x, CO, CH_4, and VOCs emissions are both natural and anthropogenic, but generally it is the regions with prevalent anthropogenic emissions that are most prone to elevated surface (lower troposphere) ozone concentrations. Due to the anthropogenic emissions of the ozone precursors, urban and developing regions usually exhibit greater surface ozone amounts than rural and remote places. Exact pathways leading to the formation of the surface ozone have been difficult to establish due to the continuously changing composition of the lower troposphere (Seinfeld and Pandis 2012).

Without any other contributors the reactions involving NO, NO_2, O, O_3, M, and radiation at wavelength < 424 nm form a null cycle:

$$NO_2 + h\nu\ (\lambda < 424\ \text{nm}) \rightarrow NO + O \qquad \text{(R1-1)}$$

$$O + O_2 + M \rightarrow O_3 + M \qquad \text{(R1-2)}$$

$$O_3 + NO \rightarrow NO_2 + O_2 \qquad \text{(R1-3)}$$

These reactions result in a steady-state, where NO_2 is created and destroyed equally. Given such chemistry, ozone concentrations would only be dependent upon the existing NO_2 concentrations; however, observations show that this is not the case. It is therefore concluded that other important chemical processes take place in the lower atmosphere (Seinfeld and Pandis 2012).

Hydroxyl (OH) radical, which forms through the reaction of excited singlet oxygen atom and water vapor molecule, has an ability to initiate oxidations of carbon monoxide (CO) and volatile organic compounds (VOCs) leading to the production of hydroperoxy radical (HO_2) and organic peroxy radicals (RO_2). The following reactions are now able to proceed:

$$HO_2 + NO \rightarrow OH + NO_2 \qquad \text{(R1-4)}$$

$$RO_2 + NO \rightarrow RO + NO_2 \qquad \text{(R1-5)}$$

It is crucial that in this case conversion from NO to NO_2 no longer requires an O_3 molecule and resulting NO_2 becomes a net source for additional O_3. In order for the ozone to grow, however, there must be a balance between NO_x and VOCs. There are possibilities for VOC-limited and NO_x-limited conditions, where the following reactions are responsible for termination of ozone production reaction chain:

VOC-limited:

$$OH + NO_2 + M \rightarrow HNO_3 + M \qquad (R1.6)$$

NO_x-limited:

$$HO_2 + HO_2 \rightarrow H_2O_2 + O_2 \qquad (R1\text{-}7)$$

$$RO_2 + HO_2 \rightarrow ROOH + O_2 \qquad (R1\text{-}8)$$

Reaction (R1-6) competes against the reaction of OH radical with generic saturated hydrocarbon (RH) that leads to the production of RO_2 needed for NO_2 formation, while the reactions (R1-7) and (R1-8) compete with the reactions (R1-4) and (R1-5) similarly impairing NO_2 production. This simplified description of surface ozone chemistry shows how an environment may significantly influence the ozone concentration at a particular location (Jenkin and Clemitshaw 2000).

Urban and developing regions are expected to see higher ozone values than rural regions as VOCs and NO_x are amply available. Rural places tend to have lesser ozone variability due to the absence of local NO_x emissions even though VOCs are still present. These cases can be called NO_x-limited. Because ozone formation depends on solar UV radiation to photolyse NO_2, the diurnal ozone cycle could be irregular for the urban places in comparison with the rural regions.

1.1.2 Air Quality Index (AQI)

In the United States, the EPA uses an air quality index (AQI) to communicate air quality forecasts to the public (Table 1). The AQI is divided into six color coded categories ranging from good (green) to hazardous (maroon). For all of the criteria pollutants (such as O_3, PM, NO_2, CO, SO_2, etc.) there is a specific way to convert individual concentrations into the AQI. Each pollutant has a capability to elevate AQI into an unhealthy range. For example, if the ozone daily maximum 8-hour average (MDA8) mixing ratio exceeds 70 ppb (137 $\mu g/m^3$) then the day would fall into the unhealthy for sensitive groups (USG) category (orange), and would count as an exceedance for a particular location. An exceedance is a violation of National Ambient Air Quality Standard

(NAAQS), which was set in 1979 by EPA in the United States to encourage reduction of air pollution. The current exceedance threshold for ozone MDA8 is 70 ppb (EPA 2015).

Table 1-1. Ozone breakpoints in ppb and AQI along with the corresponding AQ categories.

O_3 (ppb) BP_{low}-BP_{high}	AQI I_{low}-I_{high}	Category Description	Color
0-54 (8-hr)	0-50	Good	Green
55-70 (8-hr)	51-100	Moderate	Yellow
71-85 (8-hr)	101-150	USG	Orange
86-105 (8-hr)	151-200	Unhealthy	Red
106-200 (8-hr)	201-300	Very Unhealthy	Purple
201-500 (8-hr)	301-500	Hazardous	Maroon

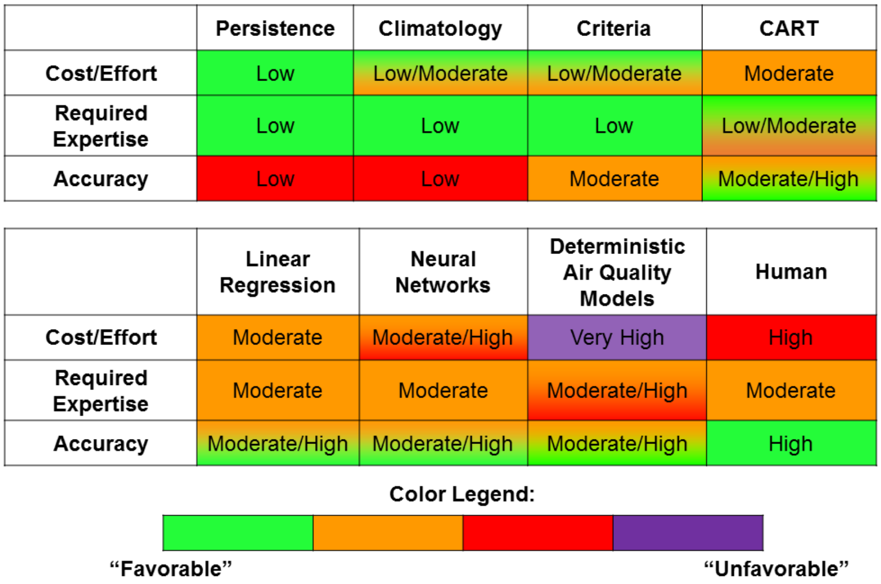
As described above, the AQI could be calculated for any of the criteria pollutants using the following equation,

$$I_p = \frac{I_{Hi} - I_{Lo}}{BP_{Hi} - BP_{Lo}}\left(C_p - BP_{Lo}\right) + I_{Lo}, \tag{1-1}$$

where I_p is the index of a pollutant p, C_p is the rounded concentration/mixing ratio of pollutant p, BP_{Hi} is the breakpoint that is greater than or equal to C_p, BP_{Lo} is the breakpoint that is less than or equal to C_p, I_{Hi} is the AQI value corresponding to BP_{Hi}, and I_{Lo} is the AQI value corresponding to BP_{Lo}. In this thesis, AQI is not used because ozone is the only pollutant examined here. However, the AQI category colorings are useful as they clearly convey the state of AQ at a given point in time. The colorings displayed in Table 1-1 are used for the MDA8 ozone forecasting products explained in Chapter 3.

1.2 Surface Ozone Prediction

To predict surface ozone a great variety of forecasting tools and methods exist (EPA 2003). Some of these approaches are relatively straightforward, while others are convoluted. Generally speaking, less elaborate ozone forecasting procedures are less accurate than more intricate ones; however, this is not always the case. Broadly, surface ozone forecasting approaches can be divided into the three groups: basic, statistical, and physically-based (Zhang et al. 2012a). Sophistication level and the cost of an approach are mostly proportional (Figure 1-1). Typically, an AQ forecaster would use an array of these approaches to arrive at the final ozone MDA8 forecast.



	Persistence	Climatology	Criteria	CART
Cost/Effort	Low	Low/Moderate	Low/Moderate	Moderate
Required Expertise	Low	Low	Low	Low/Moderate
Accuracy	Low	Low	Moderate	Moderate/High

	Linear Regression	Neural Networks	Deterministic Air Quality Models	Human
Cost/Effort	Moderate	Moderate/High	Very High	High
Required Expertise	Moderate	Moderate	Moderate/High	Moderate
Accuracy	Moderate/High	Moderate/High	Moderate/High	High

Color Legend:

🟩🟧🟥🟪

"Favorable" "Unfavorable"

Figure 1-1. Various AQ forecasting methods and their characteristics. The scale at the bottom of the figure indicates whether a property of an AQ forecasting approach is favorable or unfavorable. The challenge is to keep cost, effort and required expertise as low as possible while maintaining high accuracy (after EPA 2003).

1.2.1 Basic Approaches

Basic methods include persistence, climatology and criteria. Persistence forecasting assumes that the observed value for today is the forecast for tomorrow. The method is based on the fact that ozone significantly depends on weather conditions and that a typical weather pattern will persist for several days. The approach is fast and works well under the near-stationary and slow-moving air masses. The method does not work well when the weather regimes are rapidly fluctuating. The method's only requirement is an access to real-time and past day's ozone data. This method is commonly used as a baseline for determining if other methods have any measureable skill.

Climatology is not a standalone AQ forecasting procedure, but is used as a guide for an AQ forecaster to make sure that the forecast is reasonable. To create an ozone events climatology, historical data going back several years is required. It is imperative that the emissions during the historical record used are consistent with the current emissions. To utilize a climatology method, an AQ forecaster needs to keep a careful log of daily ozone values and compare them with climatology to see if their forecast makes sense. This approach should be used in conjunction with the other methods.

The criteria approach is based on exceeding thresholds of predicted meteorological variables that are typically observed during a high ozone day. To employ this method observed meteorological and ozone data are required as well as the corresponding forecasted meteorological data. This procedure allows a forecaster to get an idea of what to expect with regard to MDA8 ozone for a forecasted time period, but does not provide any specific values.

All of these approaches, persistence, climatology and criteria, are relatively quick and do not require high level of expertise. Consequently, the operational cost of such means is low. Their main drawback is low accuracy due to inability to handle unusual scenarios and sudden changes in weather and emissions.

1.2.2 Statistical Approaches

Examples of statistical surface ozone models include various types of least squares regressions and artificial neural networks (ANNs) methods (Prybutok et al. 2000). The main principle behind such models is a considerable correlation between weather variables and ozone. For instance, through alteration of OH radical concentrations, the moisture content in air is able to influence ozone chemistry, where wetter and cloudy conditions tend to reduce ozone while dryer conditions tend to favor ozone formation (Lelieveld and Crutzen 1990; Klonecki and Levy 1997; Murazaki and Hess 2006). Clouds may also significantly reduce photochemical processes vital for surface ozone production (Thompson 1984; Lelieveld and Crutzen 1990; Flynn et al. 2010). Air temperature is typically cited as one of the closest direct associates with surface ozone as it is able to increase rate of photolytic chemistry and also reflects the amount of surface radiation (Sillman and Samson 1995; Klonecki and Levy 1997; Aw and Kleeman 2003). Calm conditions are typically more advantageous for ozone build up as ozone is not being removed from the location where it has formed; however, ozone advection and transport from the free troposphere may present exceptions to this rule (Tu et al. 2007). In summary, there is an ample evidence that supports a strong relationship between meteorology and the formation of surface ozone, but there are many ways to define ozone dependency on meteorological variables, so many different methods are available. The most common statistical AQ approaches include classification and regression trees (CART), linear regression, extreme value approaches and artificial neural networks (Burrows et al. 1995; Van der Wal and Janssen 2000; Cobourn 2007; Shad et al. 2009).

Using various meteorological variables, the CART method generates a decision tree with a number of categories, which represent various modes of ozone pollution – some are polluted, some are moderate, and others are relatively clean. To determine critical meteorological variables correlation analysis is performed between ozone and various meteorological variables that are believed to influence ozone values the most. Once these variables are identified, they are used to split the ozone data in such a way as to produce two most dissimilar groups. Then for each group the correlation analysis is reapplied and each of these groups is split into another two groups. This process continues until the data becomes sufficiently similar and cannot be divided into distinct groups anymore. CART requires a quality assured set of ozone and meteorological data for its development. Also, it takes modest amount of knowledge and expertise in statistics to make CART operational. The advantage of such approach is that it is quick and is easy to operate. It gives a good idea of what to expect in terms of ozone pollution on a given day, but again, just like the methods discussed in the previous section, CART is unable to deal with unusual scenarios.

One of the most used methods to predict ozone is linear regression. The dependency of ozone on a meteorological variable could be described with an equation of a line or a curve, which then can be used to predict ozone. Usually, several meteorological variables and sometimes other variables, such as previous day ozone and a day of the week, are used to develop multiple linear regression equation. This approach has been widely used in AQ prediction and has shown quite a bit of success. It takes a moderate understanding of meteorology and atmospheric chemistry to develop a successful regression model, and although running the model is relatively straight forward, the output should be checked for its physical reasonableness. One of the main limitations of the regression equation method is that it tends to predict an average ozone values and does not do well with extreme values.

To complement the regression equation model, an extreme value approach can be used. It is based on the extreme value theory (Thompson et al. 2001), where a certain threshold or a hazard value is defined and probability of exceeding this value is calculated. Such model would be useful if one is particularly interested in predicting ozone exceedances (Cox and Chu 1993). Although the model is designed to identify specific scenarios it would not be able to handle unusual scenarios well because the approach still relies on the historical record.

ANNs are a powerful set of algorithms designed to learn and recognize nonlinear relationships. The concept of ANNs is inspired by biological neural networks akin to the ones that operate in a human brain. A number of disciplines have taken advantage of this versatile method (Hill et al. 1994, Gardner and Dorling 1998). In air quality forecasting, development of an ANN model requires independent sets of training, validation, and testing data. The training data is supplied into the system of nodes usually containing input, hidden, and output layers. In the training process, the system optimizes the weights of different predictor variables to reduce the error of the simulated output compared with the supplied output (part of the training data). The model uses a non-linear, sigmoid, function to adjust the weights. This is especially useful for the prediction of ozone as the ozone relationship with meteorological variables is generally non-linear. Validation data is needed to identify optimal tuning parameters during the training to avoid overfitting (Figure 1-2). And finally, testing data is needed to evaluate tuned model. However, production of such a model is costly as it takes a well-rounded understanding of ANNs to develop a proper model. There are various pitfalls that an unexperienced user may fall into (e.g. overfitting) if care is not taken when the model is being created. For that reason, the application of ANNs to real-time ozone forecasting has not been widespread. Nonetheless, once the model of this kind is made, it is easy to implement and the tool can perform well, especially during ambiguous meteorological conditions. Just like the other statistical methods, AANs rely on the historical data to make a prediction and, therefore, when the emissions are altered the performance of the method will correspondingly suffer.

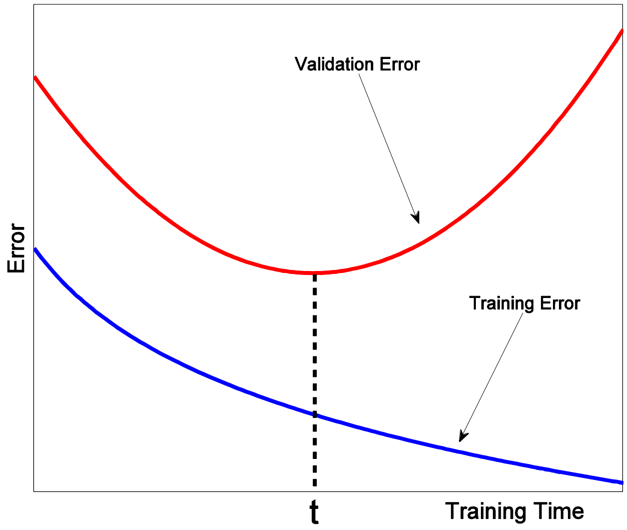


Figure 1-2. An illustration of how a validation data set is used to tune ANNs avoiding the overfitting issue. Parameter **t** on the x-axis of the plot indicates a number of iterations allowed for model optimization (after Gardner and Dorling 1998).

The advantage of the above mentioned statistical approaches is that they offer moderate to high accuracy at a moderate cost (Zhang et al. 2012a). Some of these methods have been shown to perform with significant skill and in some cases even outperformed chemical transport models (CTMs, Dutot et al. 2007). A combination of different statistical tools working together to compensate for each other's weaknesses can lead to even better predictions (Diaz-Robles et al. 2008). Although a number of statistical AQ models have been shown to perform well, most have weaknesses associated with their underlying assumptions (Thompson et al. 2001; Zhang et al. 2012a). They tend to be more accurate than basic methods, but are limited in their applications because they are confined to regions with reliable and relatively long-term ozone and meteorological measurements. Also, the nature of statistical modeling typically does not allow an understanding of physical and chemical processes that drive air pollution.

1.2.3 Physically-Based Approaches

Physically-based approaches to real-time ozone forecasting rely on deterministic atmospheric CTMs that include major meteorological, physical, and chemical interactions that govern formation and removal of ozone (Zhang et al. 2012a). To simulate ozone in such a way it is imperative to model the emissions, transport, diffusion, formation and removal of ozone and its precursors. The concept can be summarized with the following equation:

$$\frac{\partial c_{O_3}}{\partial t} + \nabla \cdot \boldsymbol{U} c_{O_3} = \nabla \rho D_{O_3} \nabla \left(\frac{c_{O_3}}{\rho} \right) + R_{O_3}\left(c_{O_3}, T, I(\lambda), etc., t\right) - S_{O_3}(\boldsymbol{x}, t), \tag{1-2}$$

where c_{O_3} is the concentration of ozone, $\boldsymbol{U}$ is the wind velocity, D_{O_3} is the molecular diffusivity of ozone, R_{O_3} is the concentration change rate of ozone through chemical reactions, which is the function of temperature T and spectral actinic flux $I(\lambda)$ among other variables, S_{O_3} is ozone sink at the location $\boldsymbol{x}$, ρ is the density of air.

In order to make an ozone prediction using the concepts described above usually three different types of models are required. These models are meteorological, emissions and chemical (Figure 1-3). Meteorological model is responsible for pollutant transport, mixing, and deposition calculations. Various meteorological fields are needed to run emissions and chemical models. Emissions models estimate the flux of various chemical species from anthropogenic and natural sources into the air for a given region. To predict ozone, it is crucial for emissions model to correctly evaluate precursors needed for ozone formation at a location of interest. And finally, the chemical model combines information from the meteorological and emission models to simulate relevant chemical reactions over time and space. Then the chemical model is able to make an ozone prediction for a given grid point and time step.

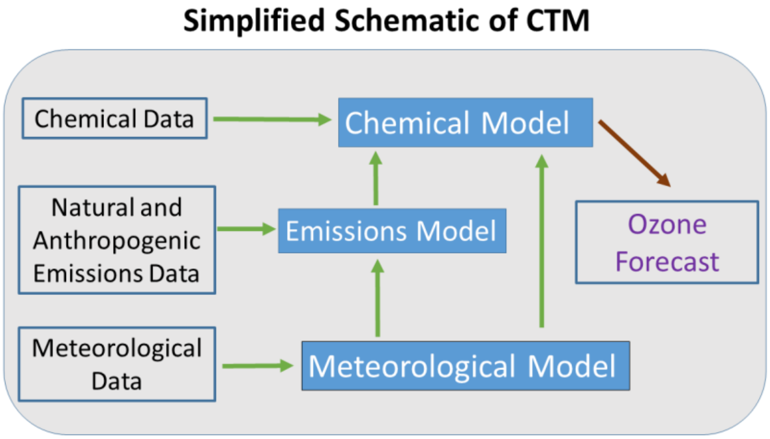


Figure 1-3. Highly simplified representation of CTM and its structure. The diagram illustrates complex synergy of many different variables.

CTMs are difficult to operate and maintain, but with a proper proficiency they usually provide forecasts with moderate to high accuracy. Additionally, they can handle well atypical scenarios such as a large biomass burning event and are spatially and temporally resolved without a need for a large quantity of measurements. The cost of running and maintaining such models is typically high.

1.3 Motivation and Goals

In 2015, the EPA reduced the MDA8 exceedance threshold from 75 ppbv to 70 ppbv across the United States (US), making the job for AQ forecasters more challenging (Cooper et al. 2015; EPA 2015). To help AQ forecasters with the prediction of ozone MDA8, various tools have been developed and evaluated over the recent decades. Despite these research efforts, only a few AQ models are available for the US AQ forecasters, the main one being the National Air Quality Forecast Capability (NAQFC, Chai et al. 2013; Stajner et al. 2014). Many AQ forecasters in the US also self-develop their own tools, sometimes based on statistical methods, to help them prepare forecasts. However, this requires considerable experience and knowledge on the part of an individual AQ forecaster (Ryan 1995; EPA 2003). The complexity of making an ozone forecast is further augmented by the fact that all of the described methods, in their typical form, are only able to make a discrete (single value) prediction meaning that the description of any uncertainty regarding a forecast is absent. Over the past decade this issue has been addressed and probabilistic approaches to AQ forecasting have been recommended (Dabberdt et al. 2004; Delle Monache et al. 2006a; Delle Monache et al. 2006b; Vautard et al. 2009, Garner and Thompson 2013).

1.3.1 Uncertainty Quantification in Ozone Modeling

The quantification of uncertainty is an important part of the forecasting trade. For instance, if a thunderstorm is in the forecast for tomorrow afternoon, how does one decide whether to cancel a sporting event or not? Of course, if the forecast is completely trusted, the event should be canceled. However, as known from experience thunderstorms are often scattered, but that is not always the case as they could be associated with a mesoscale convective system and that is why knowing the probability of a thunderstorm passing over a region becomes crucial. Although at times subjective, evaluating the probability of an event leads to better decision making, which reduces costs (Keith and Leyton 2007).

In meteorology, it is possible to produce probabilistic forecasts with the help of ensembles (Kalnay 2003). An ensemble can be comprised from different models predicting the same variable or it could be generated from the same model by varying inputs such as initial and boundary conditions. Ensembles help to quantify uncertainty and to reduce forecast error of a particular event by simulating different but plausible scenarios of this event (Delle Monache and Stull 2003). Applications of ensembles to AQ have been slow to develop as modern AQ CTMs contain numerous chemical and meteorological variables presenting computational challenges (Dabberdt and Miller 2000). Nevertheless, recent work regarding the use of ensembles in AQ forecasting is promising (Zhang et al. 2012b).

Multi-model approaches have been used in several studies (Delle Monache and Stull 2003; McKeen et al. 2005; Delle Monache et al. 2006a) showing significant improvements in reducing forecasting error with ensemble averaging. The application of multi-model ensemble is complex, however. It requires running several different CTMs simultaneously, which may not be operationally feasible. Another limitation of this approach is that the number of members typically ranges from 4 to 10 models, precluding a thorough uncertainty quantification. To generate a large ensemble of AQ predictions, a single CTM can be used by reasonably varying input variables and model parameters in order to account for intrinsic model uncertainty. This can be done with Monte Carlo (MC) simulations method, which randomly perturbs input model data based on the probability distributions chosen by the user. Although straightforward in theory, the MC simulation method is computationally demanding. Therefore, limiting perturbations to just meteorological variables presents a more realistic scenario with regard to real-time AQ forecasting. In some studies, the approaches described above have been combined with statistical post processing techniques to quantify the uncertainty of AQ forecasts. and lead to encouraging results (Zhang et al. 2012b; Garner and Thompson 2013; Djalalova et al. 2015).

Although the AQ modeling methods described above are promising, for real-time operation they may require sustained computing power currently not readily available (Carmichael et al. 2008; Wilks 2011; Zhang et al. 2012b). The alternative to using CTMs to produce surface ozone probabilistic forecasts is based on statistical AQ modeling (Zhang et al. 2012a). An AQ statistical model can greatly benefit AQ forecasters because its predictions can be readily produced for most of the AQ monitoring stations in the US, informing the forecasters the range of likely MDA8 ozone outcomes. Combining CTMs and statistical AQ probabilistic MDA8 ozone forecasts would allow

US AQ managers to make robust decisions by taking into account the forecast uncertainty (Garner and Thompson 2012, 2013).

1.3.2 Presented Approach

This thesis presents a statistical model that is not computationally as demanding as the methods described in the previous section. It encompasses a few of the mentioned statistical approaches to generate a probabilistic ozone MDA8 forecasts that quantify the uncertainty inherent to AQ prediction. The resultant hybrid statistical model combines self-organizing map (SOM) (Kohonen 2013), a type of ANN, with a stepwise weighted quadratic regression, a special case of a multiple linear regression model (Wilks 2011). SOM brings to this union an ability to distinguish synoptic-scale weather patterns (Hewitson and Crane 2002), while the linear regression is exploited to evaluate local weather effects on ozone variability. Together they capture the range of weather impacts on local ozone amounts.

The SOM algorithm performs vector quantization, a process that reduces a large set of data into a more compact representation such that the transformed data is organized with respect to its similarity, which makes it an effective tool for clustering weather maps (Pearce et al. 2011). In the presented method, different synoptic-scale weather patterns are first identified using SOMs and then a regression equation is developed for each of the patterns. The clustering of spatial meteorological data (e.g. weather maps) with SOM captures distinct synoptic patterns, where every synoptic pattern modulates surface ozone in its own unique way (Wilczak et al. 2009).

Local weather variables affecting ozone variability are not, however, exclusively controlled by the synoptic patterns. Therefore, to improve forecast accuracy, a linear regression equation is developed for each SOM-identified synoptic pattern. AQ station-specific meteorological and chemical variables are used as predictors to capture local effects.

To make a forecast, the upcoming weather pattern is predicted with a numerical weather prediction (NWP) model. The similarity of this predicted pattern to each SOM-derived weather pattern is quantified using a generalization of the Cressman weighting function (see Chapter 2). The regression equation forecasts corresponding to these SOM-derived weather patterns are combined together using weighted kernel density smoothing resulting in a probabilistic forecast of the ozone MDA8 (Wilks 2011). The calibration of this probability density function (PDF) depends upon the weights chosen for each regression and is, thus, tunable via the parameters of the generalized Cressman weighting function. This method is similar in some ways to an NWP ensemble forecasting method, as explained in section 1.3.1 and by Wilks (2011). It differs, however, in that each "ensemble member" is not driven by an equally likely NWP model run, but rather by historical weather patterns that resemble to a greater or lesser degree the pattern predicted by a deterministic NWP model (Greybush et al. 2008).

An advantage of this approach is that it does not require any input from CTMs, but instead relies solely on a deterministic NWP forecast along with a series of site-specific meteorological and ozone observational records. This forecasting method is referred to herein as regression in SOM or REGiS. Not only does REGiS allow for a classification of various ozone pollution scenarios and evaluation of their likelihood for a given deterministic weather forecast, it also provides insight into which weather patterns are favorable for ozone pollution episodes and which are not. The physical logic of the system is thus accessible to the forecaster, a trait not shared by other nonlinear probabilistic forecast tools such as ANNs. Besides serving as an additional, uncertainty quantifying guidance for US AQ forecasters, REGiS is well-suited to international AQ forecasting where a regional CTM may not be available (Frost and Meagher 2010).

The main goal of this thesis is to explain REGiS mechanics and to test the feasibility of the REGiS design. In the first part of this work (Chapters 2 and 3), it is assumed that the meteorology is known perfectly and hence reanalysis and observation data are used to generate ozone MDA8 predictions during the evaluation process. In real-time forecasting example (Chapter 4), meteorological data is supplied by NWP systems: Global Forecasting System (GFS, Environmental Modeling Center 2003) and Model Output Statistics (MOS, Carter et al. 1989) in order to run REGiS. By itself, REGiS may not provide an AQ forecaster with a precise guidance, but together with NAQFC and possibly other tools, it would be easier to make a well-informed decision regarding the predicted air quality.

In the subsequent Chapters, the foundation, implementation, tuning and performance of this statistical model are explored in detail. The model is first developed, and then evaluated using a standard set of skill metrics using ozone data from several AQ stations in the San Joaquin Valley (SJV) of California, a region well-known for poor AQ (Beaver and Palazoglu 2009). To demonstrate applicability of REGiS to various geographic regions the model is also applied to two AQ sites in northeastern Colorado (CO): long-term station and short-term DSICOVER-AQ campaign site (Crawford and Pickering 2014). The advantages and drawbacks of the proposed approach are illustrated. The brief operational period of REGiS is also discussed. Chapter 2 methodically describes the foundations and workings of the model as well as the data that is used for testing and training of the model. The operational aspects of the model and how it is tuned for optimal performance are also discussed. The results of the first part of this research are presented in Chapter 3, where the model is evaluated using several sets of independent data. In Chapter 4, a recent operational REGiS experiment is described. Chapter 5 summarizes this work and suggests the route for further investigation and application of the method.

Chapter 2

Developing Regression in SOM - REGiS

REGiS is inspired by tree-based and stratified models that are based on the idea that the association between an air pollutant and meteorology may be different in different meteorological regimes (Thompson et al. 2001). Instead of regression trees, meteorological regimes are identified by SOM (Kohonen 2013), a technique related to ANNs. Once regimes are identified, a stepwise weighted quadratic regression equation for ozone MDA8 is developed for each weather pattern. Finally, the regression forecasts are weighted, based on their SOM node's similarity to the predicted weather pattern, to obtain a PDF for ozone MDA8.

2.1 Self-Organizing Maps (SOMs)

Originally, SOMs were inspired by the problems of pattern recognition in images and speech, but over time they became used for numerous other applications such as linguistics, cryptography, and geoscience to name a few (Lindblom et al. 1983; Agarwal and Skupin 2008; Abdulkader and Roviras 2012). The ability of SOMs to extract and visualize patterns from high-dimensional data is especially useful for meteorology and there are a number of examples of this in the published literature (Hewitson and Crane 2002; Pearce et al. 2011; Jensen et al. 2012). For instance, SOMs have been helpful in studying variability of different synoptic events as well as in understanding how weather patterns change throughout the seasons in general circulation models (GCMs) (Hudson and Hewitson 2001; Hewitson and Crane 2006). An important feature of SOMs, that distinguishes it from the other clustering algorithms such as K-means or Hierarchical, is the ability to arrange a 2-D lattice of clusters according to their similarities, rather than just grouping data cases into randomly positioned categories (Agarwal and Skupin 2008). This meta-grouping is accomplished by using a neighborhood function h (shown in (2-4)), which allows clusters located next to each other in the lattice (neighbors) to influence each other's position. This component of SOMs provides a training advantage compared to the other mentioned methods because it allows each pattern to be defined by a larger sample of input data.

2.1.1 SOMs Algorithm

One way to explain the process of pattern recognition with SOMs is through a demonstration. First, 2-D data in the range [0,1] is generated. To make the example more interesting, the data is not normally distributed – the 50% of the x-coordinates of the data are sampled randomly from the [0.001,0.45] range, while the other 50% of the x-coordinates of the data are sampled from the [0.01,1] range. The dataset is visualized in the Figure 2-1. Now the goal is to identify related groupings in these data. The question is, what is the best way to cluster the points so that the distinct patterns are identified?

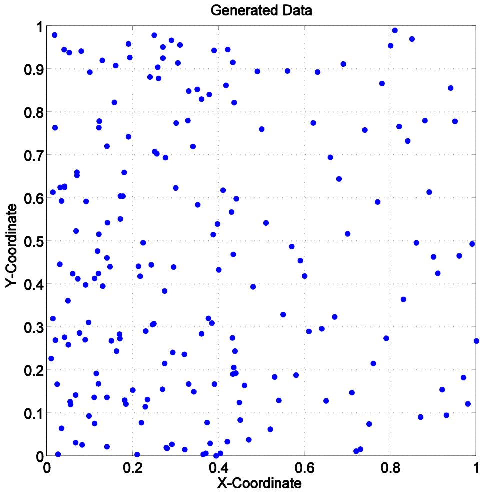


Figure 2-1. Semi-randomly generated data to demonstrate the mechanics of SOM. For more details, see text.

The SOMs algorithm begins with the initialization of the clusters, which are called nodes. The size and shape of nodes must be pre-determined. Because the results may be sensitive to the size and arrangements of nodes in the map, the dimensions of the node map (SOM_{dim}) have been one of the most frequently discussed topics in the SOMs literature (Kohonen 2013). This question does not have a "right" answer because the optimal dimensions of a SOM depend on the application. Based on a problem at hand, a compromise has to be found between resolution and statistical accuracy of the map (Kohonen 2013; Stauffer et al. 2016). In the example shown here, the size and shape of the nodes are chosen to be 9 and 3x3 respectively.

The next step is to initialize SOM nodes before the iterative training begins. Figure 2-2 shows linearly initialized nodes along the dimensions of the SOM through a uniform sampling from a subspace spanned by the first and second principal components of the input data (Johnson et al. 2008). Each node forms a Voronoi region around them, determined by Euclidian distance from a node to a point. The points in these Voronoi regions (of the nodes) are indicated with different colors. The way nodes are positioned in the Figure 2-2 is not optimal because the outliers are not well represented, which can be deduced from the visual examination. The purpose of SOMs, therefore, is to rearrange the nodes to a more meaningful configuration using an unsupervised learning algorithm.

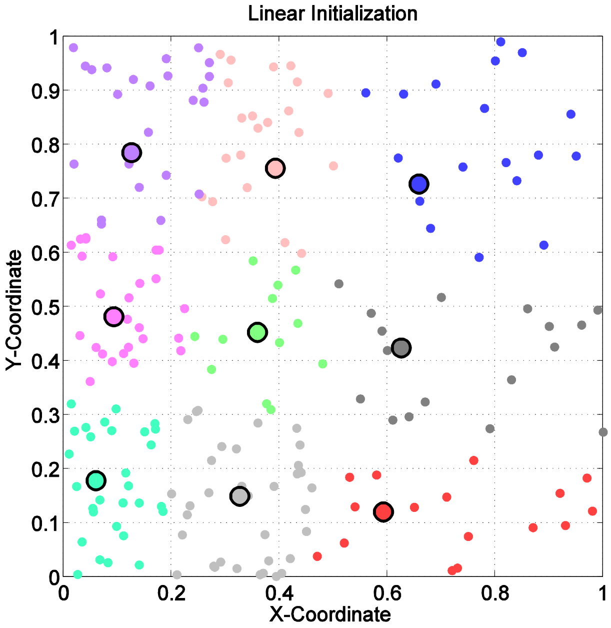


Figure 2-2. Linear initialization of the SOM applied to the data shown in the Figure 1-1.

To perform the unsupervised learning procedure with a map, batch training algorithm is used (Vesanto et al. 2000). This process is iterative, where at each step the input data are partitioned into Voronoi groups (as shown in Figure 2-2). The assignment of each data point to a particular node is based on the minimum Euclidian distance between an input point j, represented by vector $\mathbf{x}_j$, and a reference node i, represented by vector $\mathbf{m}_i$:

$$\|\mathbf{x}_j - \mathbf{m}_c\| = \min_i\{\|\mathbf{x}_j - \mathbf{m}_i\|\}, \tag{2-1}$$

where index c denotes the Best-Matching Unit (BMU), the node that is the closest to the input vector $\mathbf{x}_j$. Once all of the input data are mapped to the nodes, the nodes are updated based on a neighborhood (i.e. similarity) weighted average of the input data:

$$\mathbf{m}_{\boldsymbol{i}}(t+1) = \frac{\sum_{k=1}^{m} h_{ik}(t)\mathbf{s}_k(t)}{\sum_{k=1}^{m} n_k h_{ik}(t)}, \tag{2-2}$$

where $h_{ik}(t)$ is the neighborhood function, t is an epoch or a count of iterations, n_k is a number of training cases in node vectors in node k and $\mathbf{s}_k(t)$ is the sum of these training cases, where each training case of node k is denoted by $\mathbf{x}_{\mathrm{p}}$:

$$\mathbf{s}_k(t) = \sum_{p=1}^{n_k} \mathbf{x}_p\,. \tag{2-3}$$

The neighborhood function, $h_{ik}(t)$, measures the distance between two nodes. Its value ranges from 0 to 1, with value 1 when $i = k$ and lesser values as k moves farther away from i on the map. Different neighborhood functions are available. Here, the Epanechikov neighborhood function is chosen because it was shown to have the lowest quantization error compared to the other standard neighborhood functions that are used in SOMs (Liu et al. 2006; Stauffer et al. 2016). The Epanechikov neighborhood function has the following mathematical representation:

$$h_{ik}(t) = \max\left\{0{,}1 - \left(\frac{\|\mathrm{r}_i - \mathrm{r}_k\|}{\sigma(t)}\right)^2\right\}, \tag{2-4}$$

where $\|\mathrm{r}_i - \mathrm{r}_k\|$ is the distance between units i and k on the SOM grid and $\sigma(t)$ is the neighborhood radius at iteration t. Over the course of training, $\sigma(t)$ is decreased linearly with each repetition t until it becomes 1. The training is started with a fairly large $\sigma(t)$, at least half the diameter of the network (Kohonen 2001). In this way, a well-organized map of patterns is formed, resulting in a gradual transition from one pattern to the next as one moves across the map allowing for a smooth transition between adjacent regimes. The SOMs implementation used in this thesis is the SOM toolbox for MATLAB developed and written by the research group from the Laboratory of Information and Computer Science in the Helsinki University of Technology (available at http://www.cis.hut.fi/somtoolbox/).

Using the above-described training algorithm, the nodes, shown in Figure 2-2, are adjusted to optimally represent 200 data points with only 9 points. The process is demonstrated in Figure 2-3, where the initialization along with 10 iterations and the 100th iteration of the training are presented. Most of the training is accomplished over the first 5 iterations; however, it takes longer to get a more precise SOM representation of the data as is evident from the 100th epoch image. In particular, the dark grey node adjusts slowly to the right to improve the characterization of its Voronoi region. And comparison of the initialization step with the 100th epoch reveals that the outliers, which are blue, dark gray and red, benefit the most from the training.

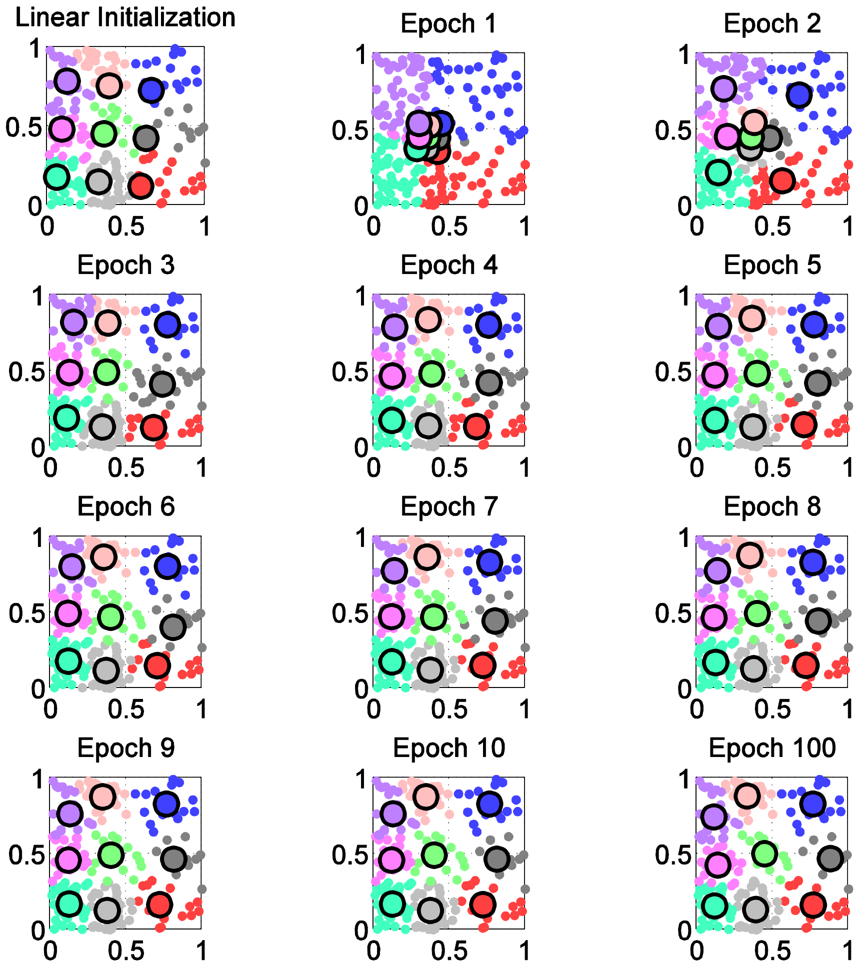


Figure 2-3. SOM training process showing linear initialization of the nodes, first 10 iterations (epochs), and the 100th iteration – fully organized map.

2.1.2 SOMs as a part of REGiS

In REGiS, SOMs serve to classify ozone-relevant synoptic-scale weather patterns over the region of interest. Similar to the example above, SOM uses pre-determined number of nodes to describe the inputted data. However, instead of 2-D points the input data consists of the vectors containing spatial weather variables such as 500 hPa geopotential height, 2-m T, 2-m T_d, 10-m wind components, etc. Because all of these variables have different magnitudes they need to be normalized for fair comparison during the SOM procedure. The SOM process is illustrated in the Figure 2-4. In steps 1 and 2, spatial weather variables are combined into an array representing the meteorological data for a single day. Steps 3 and 4 summarize the SOM training process, where the daily arrays are compared with the predetermined node arrays and the SOM map gets adjusted based on array similarity (for the details on this training process see section 2.1.1). And finally, step 5 indicates the SOM map after the training is finished.

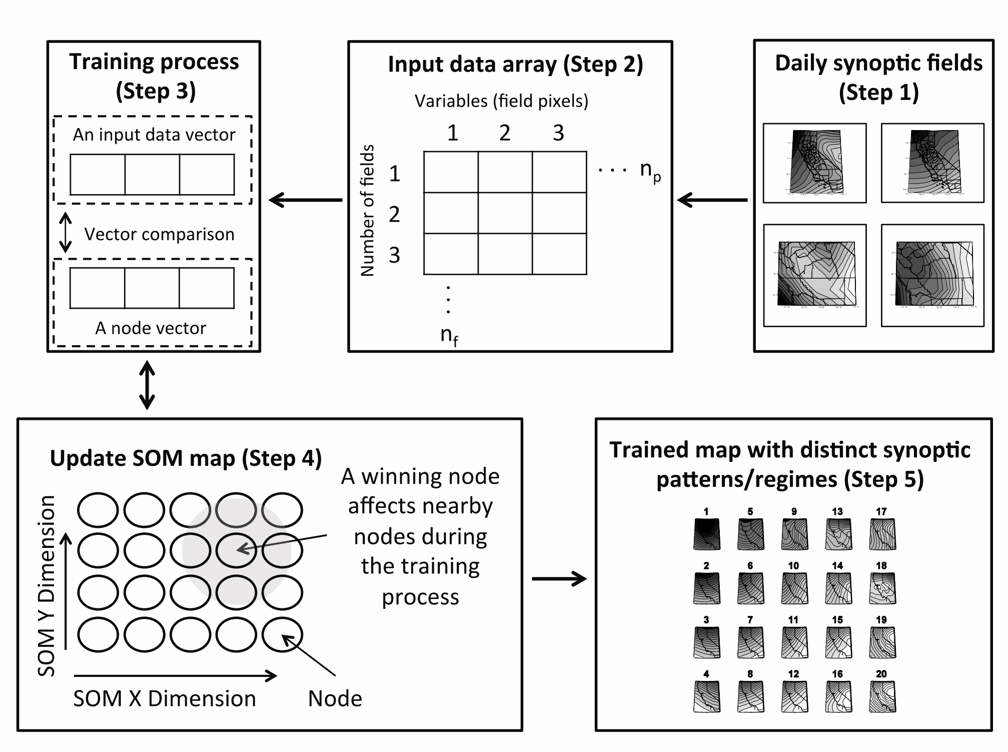


Figure 2-4. A schematic indicating major steps in the SOM algorithm to cluster daily synoptic fields into the distinct weather patterns (after Richardson et al. 2003).

Because of the high ozone pollution, the SJV in California serves as a good location to test REGiS. In order to distinguish between ozone-relevant synoptic-scale weather patterns using the

SOM method, daily 500 hPa geopotential height, 2-m T, 2-m T_d, 10-m V, 10-m U, and 850-hPa T fields on a 0.75° x 0.75° latitude/longitude grid at 0000 UTC are utilized from ERA-Interim reanalysis (Dee et al. 2011). The data is generated by the European Centre for Medium-Range Weather Forecasts (ECMWF) and is available free of charge at the following website: http://apps.ecmwf.int/datasets/. These variables are selected because they capture transport (500 hPa height and 10 m wind components) and stability (500 hPa height, 850 hPa temperature, 2-m temperature and 2-m dewpoint), which primarily provide links between synoptic-scale weather and local ozone concentration. To capture these local effects, it is imperative to define appropriately the domain over which the synoptic patterns are determined. The domains used in the first part of this study are centered on the states of CA and CO (Figure 2-5). The second part of the study looks at AQ in Philadelphia region and uses corresponding domains (for more details see Chapter 4). Because 500 hPa can be thought of as a steering level (Carlson 1991), meaning that the synoptic systems close to the surface move in response to the larger-scale wind patterns at 500 hPa, the domains for 500-hPa geopotential height fields (Figures 2-5a and 2-5c) are slightly larger than that for the other 5 variables (indicated by green rectangles in Figure 2-5). The SOM of the mentioned variables for June-July-August (JJA) of 1987-2012 is shown in Figure 2-6.

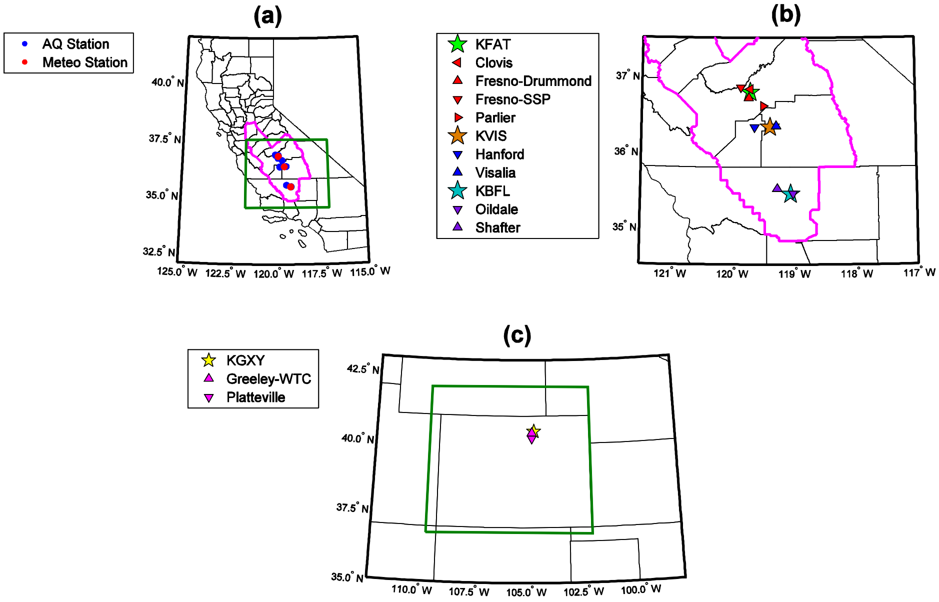


Figure 2-5. Domains used by SOM when identifying synoptic patterns for REGiS with AQ and meteorological stations relevant to this study. (a) Domain over SJV (outlined by the purple contour) that is used for 500 hPa geopotential height fields. (b) Domain (also indicated by the green rectangle in (a)) used for 2-m T, 2-m Td, 10-m V, 10-m U and 850 hPa T fields. (c) Domains as described in (a) and (b), but located over CO.

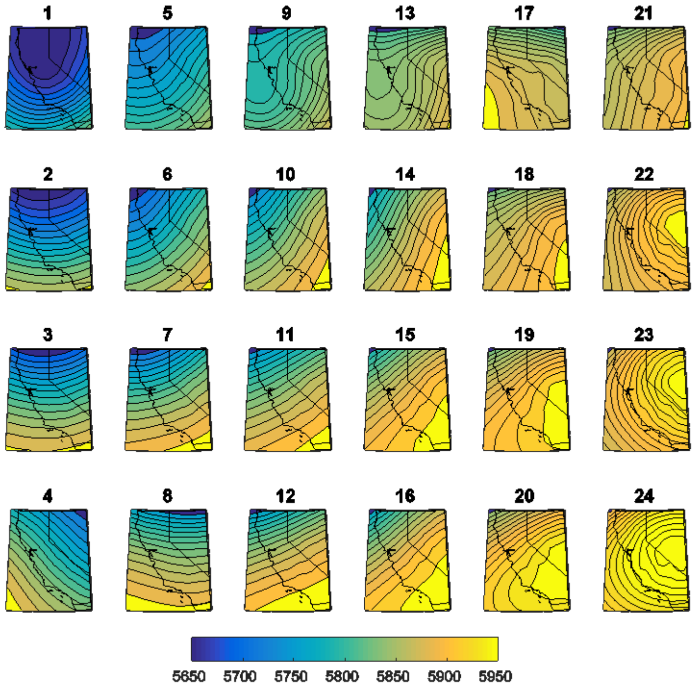


Figure 2-6. 4x6 SOM of daily 500 hPa geopotential height, 2-m T, 2-m T_d, 10-m V, 10-m U, and 850 hPa T fields from JJA over 1987-2012, centered on the Western United States (domain shown in Figure 2-5). Note: only 500 hPa geopotential height fields (in meters) are shown in the full SOM. The data used for this analysis come from ERA-Interim reanalysis.

2.2 Regression

Once the distinct synoptic regimes are established via SOMs (Figure 2-6), a stepwise weighted quadratic regression equation is developed for each regime using local meteorological and chemical variables as predictors for ozone MDA8 (Comrie 1997). The regression method used in REGiS is akin to the one described in section 1.2.2. The difference is that REGiS contains as many regression equations as there are SOM nodes. The idea is that each equation will correspond to a particular meteorological setting, better capturing ozone-meteorology sensitivity given a specific weather regime. When making a forecast, REGiS will typically use most of the equations, but the higher weights will be given to the equations representing weather patterns that closer resemble the predicted pattern.

2.2.1 Predictand and Predictors

The production of surface ozone is significantly affected by meteorology as has been discussed in section 1.2.2. This fact allows for ozone MDA8 prediction using causal meteorological variables shown in Table 2-1. Surface temperature affects the rate of chemical reactions and considered to be one of the most important meteorological predictors of ozone, where higher temperatures tend to lead to higher ozone concentrations (Sillman and Samson 1995). Dew point temperature predictor is used as a proxy for atmospheric moisture content that alters hydroxyl radical concentrations, where wetter conditions tend to reduce ozone while dryer conditions tend to increase ozone formation (Lelieveld and Crutzen 1990). Another commonly used pair of predictors for ozone are wind direction and speed (Tu et al. 2007). Certain wind directions are more favorable for the development of a pollution episode because they carry ozone precursors from the emission-heavy regions. Often, light wind speeds are indicative of stagnation as they allow for the pollutants to accumulate in the atmospheric boundary layer. Because ozone formation is dependent on photochemistry, cloud cover plays an important role in regulating ozone and is frequently used as a predictor (Thompson 1984). The two additional predictors employed by REGiS are zenith angle and the previous day ozone.

Table 2-1. Predictors used in REGiS. Analysis (not shown here) indicates that the ozone MDA8 period most often occurs either from 1000-1800 or from 1100-1900 local time.

Variable	**Time Scale**	**Units**
Zenith angle	Daily at 1200	degrees

Surface temperature	Mean over 1000–1800	°C
Dew point temperature	Mean over 1000–1800	°C
Wind direction	Mean over 1000–1800	degrees
Wind speed	Mean over 1000–1800	meters/second
Sky cover	Mean over 1000–1800	oktas
Previous day ozone	Previous day MDA8	ppbv

The reason for the inclusion of zenith angle is that the angle at which solar radiation strikes the surface of the Earth influences the photolysis process that is responsible for generation of ozone (Seinfeld and Pandis 2012). In summer, the zenith angle is smaller and more energy is available for NO_2 photolysis, but in fall and spring zenith angle is larger and less energy reaches the surface of the Earth. The zenith angle is potentially an important predictor when the regression training period occurs between summer and fall or between spring and summer. Although a day in July and September may experience similar high temperatures the photolysis rate usually differs and correspondingly affects ozone production.

The previous day ozone MDA8 is a powerful predictor because of the episodic nature of ozone pollution (EPA 2003). Sometimes an AQ forecaster would use previous day ozone MDA8 as a next day ozone forecast. Such process is known as persistence forecasting. Adding previous day ozone to the multiple linear regression model usually increases the skill of the model, but sometimes this may lead to a large error. The described predictor tends to influence ozone prediction significantly. To weaken the effect of the previous day ozone predictor on the final forecast it is possible to request REGiS to perform two regressions: one with the previous day ozone and another without. For instance, if SOM identified 24 patterns then there would be 48 regression equations – two equations for each pattern. In this way, possible ozone MDA8 scenarios are better represented. The option with the two regression equations per regime is used in this thesis.

The data to fit the regression equations come from an AQ station and the nearest meteorological station. Figure 2-5 shows the map of all the stations that are used to evaluate REGiS. Additionally, Table 2-2 summarizes the AQ and correpsonding meteorological stations used in the REGiS evaluation process. Fresno Air Terminal (KFAT) data are used for Clovis, Fresno-Drummond, Fresno-SSP and Parlier. Visalia Municipal Airport (KVIS) data are used for Hanford and Visalia – N. Church. Meadows Field Airport (KBFL) is used for Oildale and Shafter. Greeley-Weld County Airport (KGXY) is used for Greeley – Weld County Tower (WTC) and Platteville. The air quality data are acquired from the EPA's Air Quality System (AQS) database (http://www.epa.gov/ttn/airs/airsaqs/) and the meteorological data are downloaded from the National Oceanographic and Atmospheric Administration's (NOAA) National Climatic Data Center (NCDC) (https://www.ncdc.noaa.gov/).

Table 2-2. AQ stations used to evaluate REGiS and correpsonding meteorological (METEO) stations along with the related basic information.

Station Name	**WBAN/AQS ID**	**Type**	**Latitude**	**Longitude**	**Elevation (m)**
Fresno Air Terminal (KFAT)	93193	MET	36.78°	-119.72°	100
Clovis	06-019-5001	AQ	36.82°	-119.72°	86
Fresno-Drummond	06-019-0007	AQ	36.70°	-119.74°	89
Fresno-SSP	06-019-0242	AQ	36.84°	-119.87°	65
Parlier	06-019-4001	AQ	36.60°	-119.50°	78
Visalia Municipal Airport (KVIS)	93144	MET	36.32°	-119.39°	90

Hanford	06-031-1004	AQ	36.31°	-119.64°	99
Visalia – N. Church	06-107-2002	AQ	36.33°	-119.29°	97
Meadows Field Airport (KBFL)	23155	MET	35.43°	-119.05°	151
Oildale[*]	06-029-0232	AQ	35.44°	-119.02°	180
Shafter[**]	06-029-6001	AQ	35.50°	-119.27°	126
KGXY	24051	MET	40.44°	-104.63°	1432
Greeley – WTC	08-123-0009	AQ	40.39°	-104.74°	1484
Platteville	N/A	AQ	40.18°	-104.73°	1522

*Full name of Oildale AQ station is 3311 Manor St., Oildale.

**Full name of Shafter AQ station is 548 Walker St., Shafter, CA., 93263.

2.2.2 Detrending Ozone MDA8 Data

In order to partially account for the changes in ozone precursors emissions over the years (Jhun et al. 2015), MDA8 ozone data are detrended after the data have been separated by the SOM (Figure 2-7). The process of data detrending occurs before it is used by REGiS for the regression model development. First, a best-fit line is generated for a given MDA8 ozone time series using the least-squares method. Detrending is performed by subtracting the mean of the best-fit line from the MDA8 ozone data (Figure 2-8a). Once the detrended data is calculated the entire detrended MDA8 time series are adjusted up by the difference between the last value in the original MDA8 ozone data and the last value in the detrended MDA8 ozone data to generate the final detrended data (Figure 2-8b).

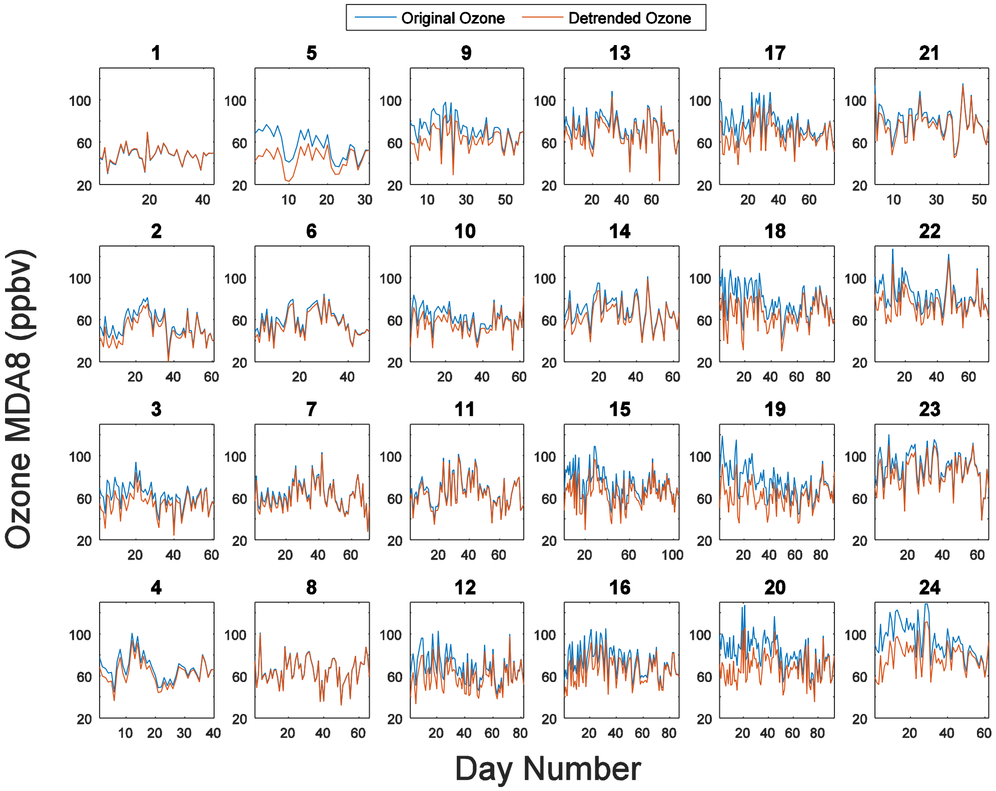


Figure 2-7. Ozone MDA8 at AQ station Fresno-SSP corresponding to 4x6 SOM over 1995-2012. Blue graph shows original ozone data and red graph is detrended ozone. Figure 2-8 illustrates the process of detrending for pattern 19.

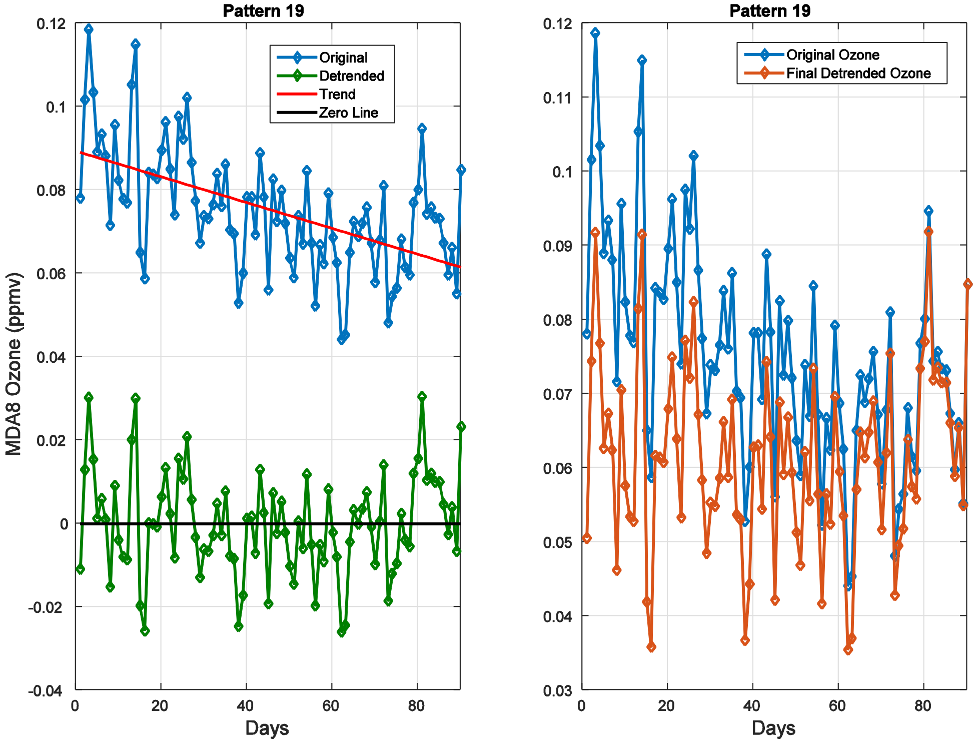


Figure 2-8. (a) Process of detrending 1995-2012 ozone MDA8 data at AQ station Fresno-SSP for pattern 19 identified by SOM method (see Figure 2-7). Original MDA8 ozone data is shown in blue with the red linear trend line. Green indicates detrended ozone. **(b)** Comparison of the original MDA8 ozone (blue) and final detrended data used in REGiS (red).

2.2.3 Regression Model Properties

Once the data is detrended for each SOM pattern, the regression equations are developed for the corresponding data. Note that the use of too many predictors in a regression equation is not recommended as this may lead to an unstable result when applied to independent data (Wilks 2011), i.e. over fitting. Therefore, a forward and backward stepwise regression is performed here, where the terms are added to or removed from a regression model based on the p-value for an F-test of the change in the sum of squared errors (Wilks 2011). The regression model is made quadratic rather than linear to allow for the nonlinearity of the meteorology-ozone relationship (Comrie 1997).

When a regression is developed for a specific regime, training data from all the other regimes may also be used, with weights based on their similarity to this specific regime. The weights ω_{ji}, of the data at nodes j, for the regression centered on the node i are calculated as follows:

$$\omega_{ji} = \left[\frac{\left(\frac{1}{d_{ji}}\right)^{\alpha}}{\sum_{j=1}^{m} \left(\frac{1}{d_{ji}}\right)^{\alpha}} \right]^{\beta} h_{ij}, \tag{2-5}$$

where α and β are the tuning parameters, d_{ji} is the Euclidean distance between the node centroids and h_{ij} is the neighborhood function defined in (2-4). The neighborhood function is used here to make the distribution of the node weights consistent with the initial SOM training process. The neighborhood radius σ determines the extent to which training cases from the surrounding nodes are used. When $\alpha \rightarrow \infty$ and $\beta \rightarrow \infty$ the weights spread out more evenly among the nodes. This process allows for a distinct model for each node (i.e. synoptic weather pattern) while still taking into account relevant information from all training cases.

2.3 Tuning and Verification

Before REGiS is used for operational forecasting, it must be trained, tuned and tested on historical data (Gardner and Dorling 1998). The training and tuning stages are performed with the historical reanalysis and observational data. In this study, the tuning of REGiS is undertaken at a single AQ site in the SJV (Figure 2-5) – Parlier – using four different training periods during June-July-August (JJA): 1987-2012, 1995-2012, 2000-2012 and 2005-2012. REGiS is uniquely tuned for each of the mentioned training periods using JJA 2013-2014 validation data set. The use of the multiple training periods is motivated by the change in emissions over time (McDonald et al. 2012). Only JJA is addressed in this work as ozone MDA8 tends to maximize in US over this time period.

To maintain independence, the Parlier site is not included in the final testing of REGiS performance. Subsequent REGiS training for the independent AQ sites is conducted using the configuration determined from Parlier and the nearby meteorological site KFAT. In the testing stage, forecasts for these AQ sites are made for periods independent from their respective training periods, which vary from station to station. The results of these tests are evaluated in Chapter 3.

2.3.1 Forecasting Procedure

In order to produce probabilistic ozone MDA8 forecasts, REGiS combines regression forecasts from the SOM nodes that have weather patterns similar to the predicted pattern (predicted by NWP model) for a time of interest. The steps in this process are thus: identification of the forecast weather pattern, assessment of its similarity to each of the SOM nodes, and combination of the regression forecasts from these nodes to produce an ozone PDF forecast. The process is summarized in Figure 2-9.

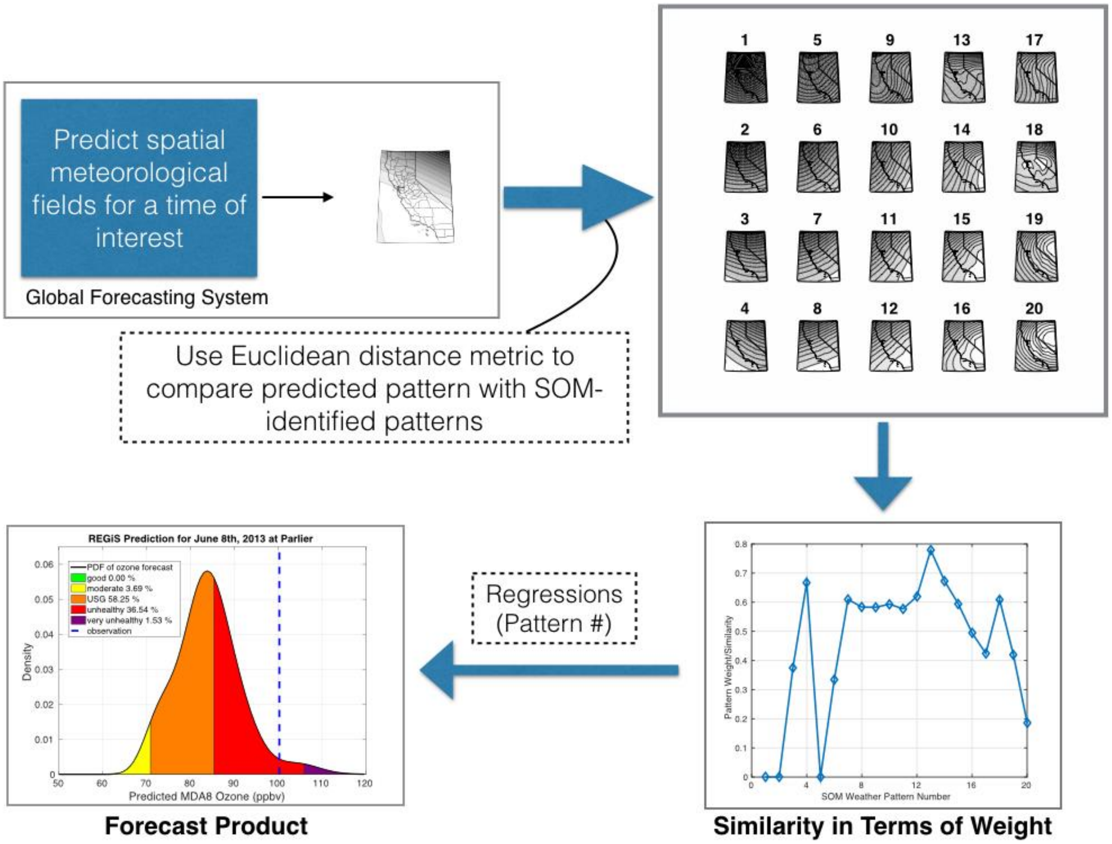

Predict spatial meteorological fields for a time of interest
Global Forecasting System
Use Euclidean distance metric to compare predicted pattern with SOM-identified patterns
1
5
9
13
17
2
6
10
14
18
3
7
11
15
19
4
8
12
16
20
Regressions (Pattern #)
REGiS Prediction for June 8th, 2013 at Parlier
PDF of ozone forecast
good 0.00 %
moderate 3.69 %
USG 58.25 %
unhealthy 36.54 %
very unhealthy 1.53 %
observation
Density
Predicted MDA8 Ozone (ppbv)
0
0.01
0.02
0.03
0.04
0.05
0.06
50
60
70
80
90
100
110
120
Forecast Product
Pattern Weight/Similarity
SOM Weather Pattern Number
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0
4
8
12
16
20
Similarity in Terms of Weight

Figure 2-9. The REGiS forecasting procedure schematic showing the steps that the model takes to produce a probabilistic forecast. First, the synoptic variables are extracted from a numerical weather prediction model (such as the Global Forecasting System) for the desired forecast time. These synoptic variables are compared with SOM-identified synoptic patterns using a Euclidean distance metric. Then the similarity between predicted pattern and identified patterns is presented in terms of weights. The higher the weight of a SOM pattern, the more similar it is to the predicted pattern. Finally, the regression equations of each pattern are solved and combined according to the estimated pattern weights to produce continuous probability density function (using kernel density estimation) of the ozone MDA8 forecast.

Recall that the weather pattern (i.e. SOM node) is determined based on the spatial distribution of several meteorological variables on the two domains (see section 2.1.2). For the prediction these fields are extracted from the NWP (such as GFS) deterministic forecast for the desired lead time.

This predicted weather pattern is then compared with all of the patterns identified by SOM (for the example of identified SOM patterns see Figure 2-6) using Euclidean distance metric d.

Because some forecast-to-node distances d are smaller and others are greater, it is possible to assign a similarity weight W between a predicted pattern and a SOM pattern. This is accomplished by using a generalization of the Cressman weighting function (Cressman 1959)

$$W = \left(\frac{R_c^2 - d^2}{R_c^2 + d^2}\right)^{\gamma}, \tag{2-6}$$

where R_c is an adjustable radius of influence around the predicted pattern and γ is a tuning parameter (Figure 2-10). Note: W is set to zero if $R_c > d$. In other words, R_c determines the cutoff threshold beyond which the regression equation of a SOM-pattern is not considered in making a prediction of the ozone PDF. Fine-tuning R_c and γ allows for various weight configurations, which influence the forecast probability distributions of MDA8 ozone. These two parameters must, therefore, be tuned in order for REGiS to produce reasonably well calibrated probabilistic forecasts of ozone MDA8.

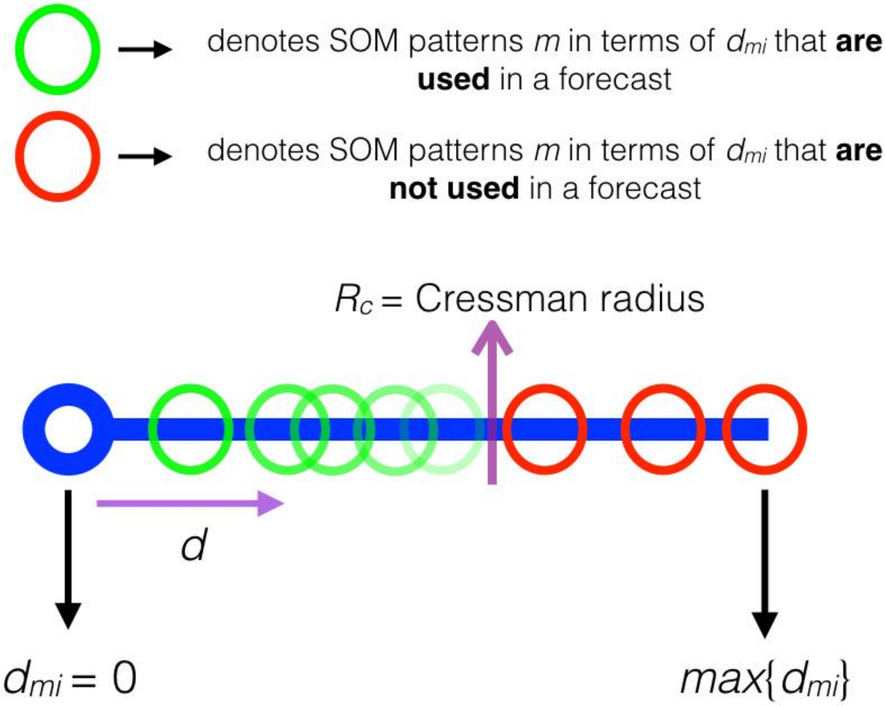

denotes SOM patterns m in terms of d_{mi} that **are used** in a forecast
denotes SOM patterns m in terms of d_{mi} that **are not used** in a forecast
R_c = Cressman radius
d
$d_{mi} = 0$
$max\{d_{mi}\}$

Figure 2-10. Diagram illustrating Cressman radius R_c, which is set by a user. R_c is the influence radius of the predicted weather pattern i on the weight of SOM weather patterns m (nodes). It takes a value of some Euclidian distance d_{mi} (distance between patterns i and m) from $d_{mi} = 0$ to $max\{d_{mi}\}$.

These regression forecasts and their weights, therefore, are used as input into the kernel density smoothing function (Wilks 2011) to produce a PDF for ozone MDA8:

$$\hat{f}(x_0) = \frac{\sum_{i=1}^{n} K\left(\frac{x_0 - W_i x_i}{h}\right)}{nh\sum_{i=1}^{n} W_i}, \tag{2-7}$$

where $\hat{f}(x_0)$ is the probability density as a function of x_0, which represents all ozone MDA8 possibilities for a given probability density function, n is the number of forecasts, W_i is a weight of a regression forecast x_i, K is a smoothing kernel function, and h is a bandwidth. Kernel density smoothing is accomplished by stacking modeled kernel shapes at each of the available values. For

K, a Gaussian smoothing kernel is used here because in this way the distribution tails are not cut abruptly so more extreme scenarios are considered. The kernel density is sensitive to h, which is estimated using an optimizing algorithm described in Bowman and Azzalini (1997).

2.3.2 Tuning

Tuning REGiS consists of finding optimal values for the parameters α, β, σ, γ, R_c and SOM_{dim} in the above-described model using a data set independent of the one used to train the SOM and the regression models. The goal here is to ensure that the PDF forecast represents a true distribution of a predicted variable. The verification rank histogram (Hamill 2001) is used for this purpose. Note that while rank histograms have frequently been applied to ensemble forecasts, the approach is equally applicable to PDF forecasts. To construct such a rank histogram, the PDF of a forecast is divided into $N+1$ segments of equal probability. This process creates $N+1$ categories into which the verifying observation may fall. The predicted PDF is statistically indistinguishable from the verifying observation if these observations fall with an equal probability in each of those $N+1$ categories. Thus, in order for the REGiS prediction system to be reliable, i.e. produce a well-calibrated PDF, the corresponding rank histogram must be relatively flat. The departure from flatness of a rank histogram can be estimated by

$$RMSD = \sqrt{\frac{1}{N+1}\sum_{k=1}^{N+1}\left(s_k - \frac{M}{N+1}\right)^2}\,, \tag{2-8}$$

where RMSD is the root-mean-square deviation from complete rank histogram flatness, $N+1$ is the number of equal-percentile regions, M is the total number of observations, and s_k is the number of observations in each particular percentile region (Wilson et al. 2007). The smaller a RMSD is, the flatter is a corresponding rank histogram. The idea is to tune REGiS parameters in such a way as to keep the verification rank histogram as flat as possible.

The objective of the tuning process is to minimize RMSD to ensure that the probabilistic forecasts are reliable. The tuning of the probabilistic REGiS is sequential. First, values for α, β, σ, γ, R_c and SOM_{dim} are assigned by an educated guess. Then one of the parameters is iterated over, while the other five parameters are held constant. The optimal value for each of the parameters is based on minimizing RMSD. The example of the tuning process at Parlier for the time period of JJA 1987-2012 is shown in Figure 2-11. RMSD is determined using the validation data over JJA 2013-2014. To begin, an educated guess is made for the six parameters. Then for $\sigma = 2$ and $\sigma = 3$, α and β are tested. For $\sigma = 1$, α and β make no impact. Lowest RMSD occurs when $\sigma = 2$, $\alpha = 0$ and $\beta = 0.5$. Next γ is adjusted, while all of the other parameters are held constant. Minimum RMSD happens when $\gamma = 0$. Using the same approach R_c is determined to be 100. Finally, iteration over different SOM_{dim} is performed.

Four tuning processes are completed in this work using the procedure described above for four different time periods using the Parlier data. Figure 2-12 shows the results of the final tuning step. The color plot shows RMSDs for probabilistic daily REGiS forecasts (for JJA 2013-2014) given various SOM_{dim}. The color bar represents RMSD in terms of cases (observation points). Figure 2-13 presents 10 of the best REGiS SOM configurations from Figure 2-12 sorted by RMSD. Results indicate that in theory, using training periods JJA 1995-2012 and JJA 2000-2012 should yield better REGiS performance than when using the other two periods.

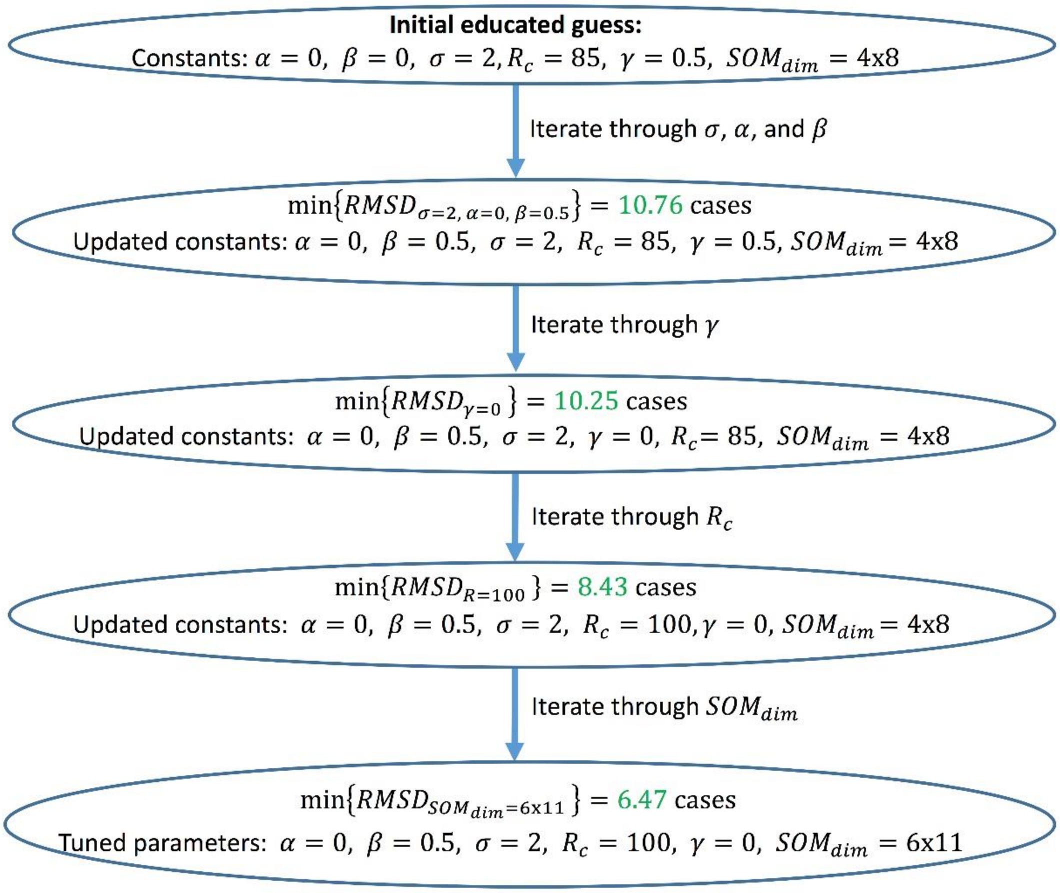

Initial educated guess:
Constants: $\alpha = 0,\ \beta = 0,\ \sigma = 2, R_c = 85,\ \gamma = 0.5,\ SOM_{dim} = 4\text{x}8$
Iterate through σ, α, and β
$\min\{RMSD_{\sigma=2,\ \alpha=0,\ \beta=0.5}\} = 10.76$ cases
Updated constants: $\alpha = 0,\ \beta = 0.5,\ \sigma = 2,\ R_c = 85,\ \gamma = 0.5, SOM_{dim} = 4\text{x}8$
Iterate through γ
$\min\{RMSD_{\gamma=0}\} = 10.25$ cases
Updated constants: $\alpha = 0,\ \beta = 0.5,\ \sigma = 2,\ \gamma = 0,\ R_c = 85,\ SOM_{dim} = 4\text{x}8$
Iterate through R_c
$\min\{RMSD_{R=100}\} = 8.43$ cases
Updated constants: $\alpha = 0,\ \beta = 0.5,\ \sigma = 2,\ R_c = 100, \gamma = 0,\ SOM_{dim} = 4\text{x}8$
Iterate through SOM_{dim}
$\min\{RMSD_{SOM_{dim}=6\text{x}11}\} = 6.47$ cases
Tuned parameters: $\alpha = 0,\ \beta = 0.5,\ \sigma = 2,\ R_c = 100,\ \gamma = 0,\ SOM_{dim} = 6\text{x}11$

Figure 2-11. Example of the REGiS tuning process at Parlier AQ station. In the given example, the training data period is JJA 1987-2012 and the validation data period is JJA 2013-2014.

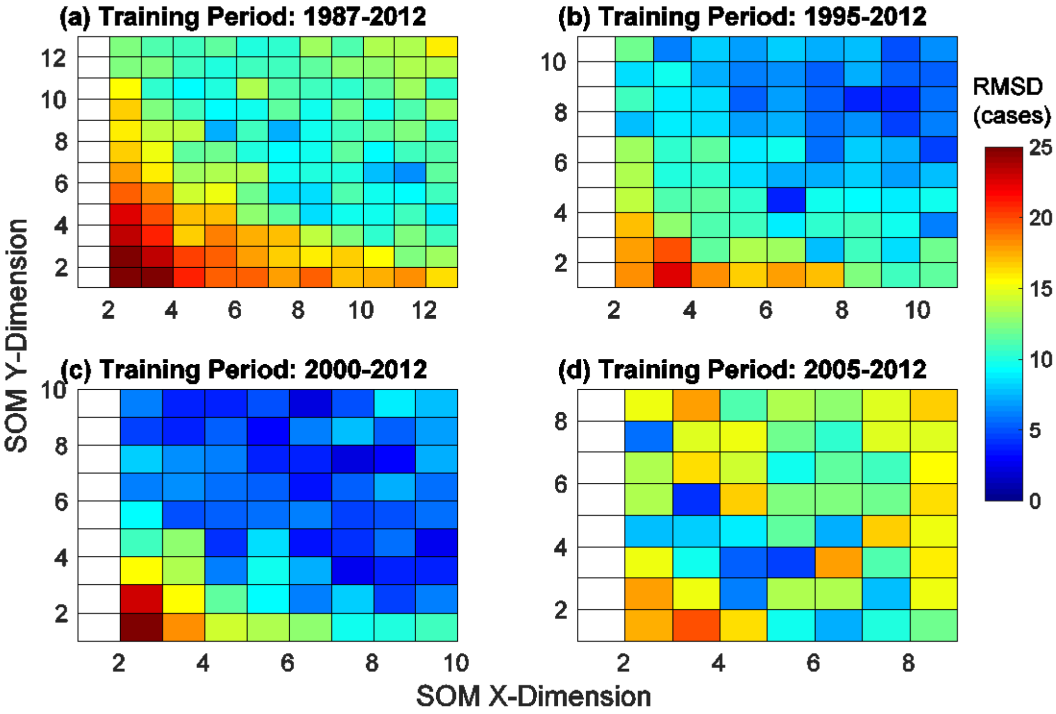


Figure 2-12. REGiS tuning process where α, β, σ, γ and R_c parameters are held constant while iterations over different and evaluations of the following RMSD values are performed. The tuning is accomplished over four different time periods: (a) 1987-2012, (b) 1995-2012, (c) 2000-2012 and (d) 2005-2012. Note that the values for the first column are not computed.

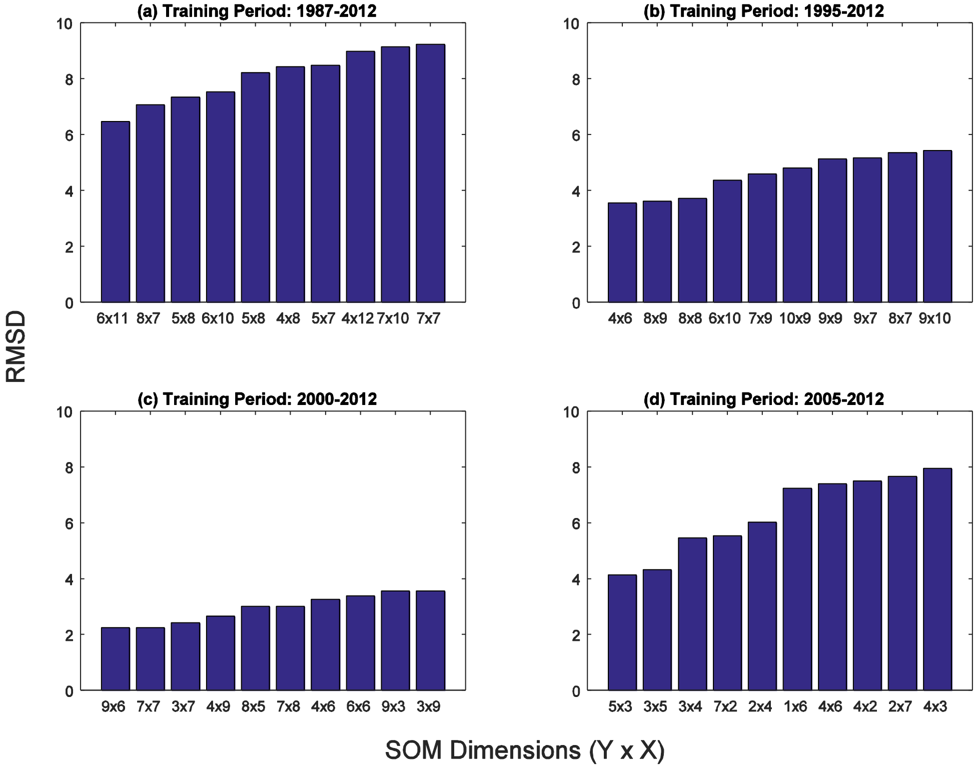


Figure 2-13. Sorted 10 minimum RMSD values from Figure 2-15. The tuning is performed over four different time periods: **(a)** 1987-2012, **(b)** 1995-2012, **(c)** 2000-2012 and **(d)** 2005-2012.

2.3.2 Verification

While rank histograms and the associated RMSD can determine if a probabilistic forecast is well calibrated, it does not quantify forecast accuracy. This latter role is filled by continuous rank probability score (CRPS) which is designed to evaluate reliability, resolution and uncertainty of a probabilistic forecast for which $F(y)$ is the cumulative probability function defined over predictand y:

$$CRPS = \int_{-\infty}^{\infty} [F(y) - F_0(y)]^2 dy, \tag{2-9a}$$

where

$$F_0(y) = \begin{cases} 0, y < observed\ value \\ 1, y \geq observed\ value \end{cases} \tag{2-9b}$$

is a step function that switches from 0 to 1 at the observation point (Hersbach 2000). Lower CRPS indicates a better model performance for a given set of data. CRPS rewards cases where an observation falls closer to the mean of predicted PDF. In this work, CRPS is used as an additional skill metric when comparing REGiS performance among different stations.

Chapter 3

REGiS Evaluation

When the REGiS completes its forecasting process, the product in a form of PDF is generated. This product is essentially a probabilistic forecast. The goal of this Chapter is to evaluate this type of forecasts. REGiS is tested using the 2013-2014 JJA data from 6 different stations independent from Parlier. The model settings used, however, are tuned using Parlier as explained in the Chapter 2.

3.1 The Product

Figure 3-1 shows a sample forecast produced by REGiS – the PDF of ozone MDA8 forecast for the AQ station Fresno-Drummond (FD) on June 11th, 2014. In the figure, ozone MDA8 runs along the x-axis and the probability density is along the y-axis. The black line indicates PDF, the area of which adds up to one. Shaded areas inside the PDF are the color-coded ozone pollution air quality index (AQI) categories, converted to ozone mixing ratios (ppbv) from the AQI (see Chapter 1) as designated by EPA (EPA 2006, EPA 2015), where green, yellow, orange and red indicate good (0-54 ppbv), moderate (55-70 ppbv), unhealthy for sensitive groups (USG) (71-85 ppbv) and unhealthy (86-105 ppbv) AQI categories respectively. The probability of each AQI category occurring for a predicted day is indicated in the legend located in the upper left corner. In the given example, there is a probability of about 87% that the ozone MDA8 exceedance (ozone MDA8 above 70 ppbv) will not occur, and indeed, on the mentioned day the observed ozone MDA8 is about 67 ppbv (blue dashed line), falling in the upper end of moderate category.

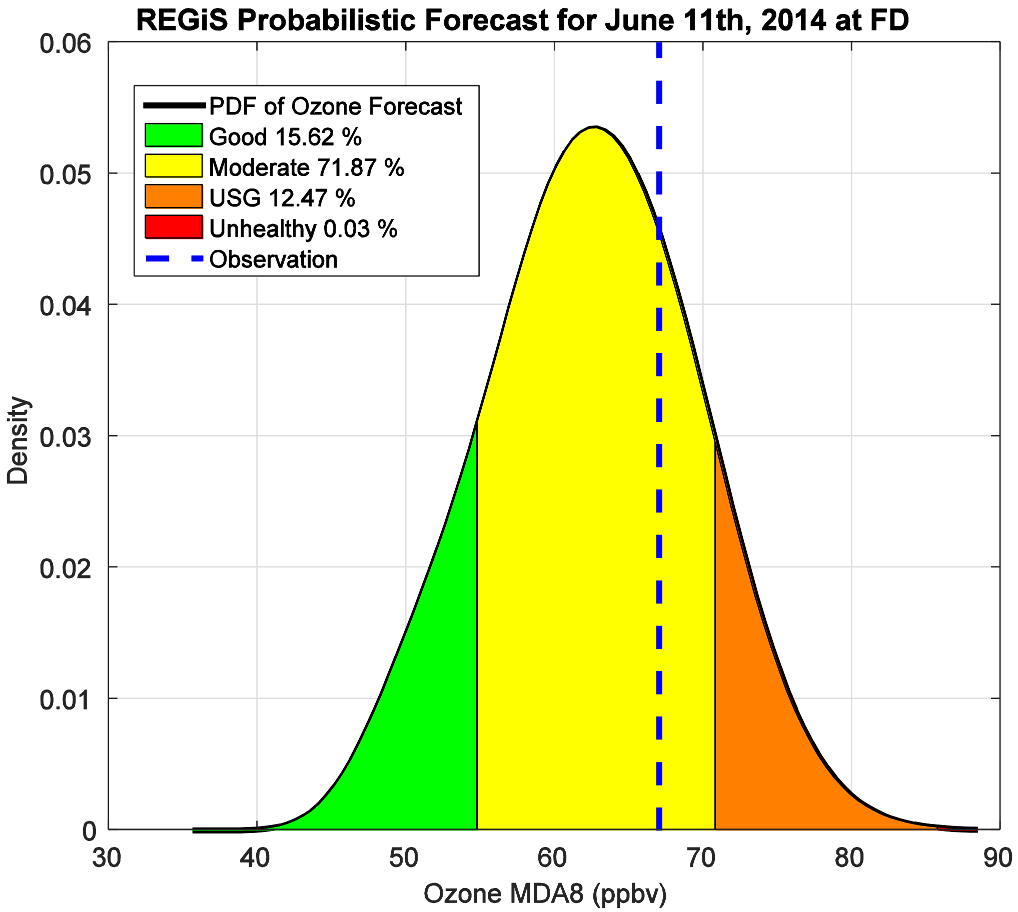


Figure 3-1. An example of the REGiS product providing probabilistic ozone MDA8 forecast at FD AQ station for June 11th, 2014.

3.2 Verification Results

To evaluate REGiS, the procedure described above has been carried out for the summers (JJA) of 2013-2014 at AQ sites in the SJV and northeastern plains of CO (Table 3-1). This evaluation period is independent from that used for training (see Chapter 2 for more details). Different training periods are tested at each station, while the tuning parameters determined in the Chapter 2 for Parlier are used. Various training periods are examined because of constantly changing ozone precursor emissions at each site. Although in this study the ozone MDA8 time series are linearly detrended before they are used in the development of the regression equations, many of the emission changes are not linear and may not be well captured with the assumption of a linearly decreasing trend. Choosing a shorter training period allows for more relevant ozone precursor emissions to be considered in the prediction regression equations. For this reason, REGiS is tested four times, decreasing its training period every time in order to understand the dependency of REGiS performance on the length of the training period. To evaluate how REGiS does at each of the stations, skill metrics of CRPS and RMSD are shown in the 5th and 6th columns of Table 3.

Table 3-1. Summary of the REGiS evaluation at 9 different stations over the SJV and northeastern CO.

Station	**SOM$_{\text{dim}}$**	**Training Period**	**Evaluation Period**	**CRPS (ppbv)**	**RMSD (cases)**
Clovis **(a)**	4 x 6	1995-2012 JJA	2013-2014 JJA	4.95	12.0
Clovis **(b)**	7 x 7	2000-2012 JJA	2013-2014 JJA	4.71	10.4
Clovis **(c)**	5 x 3	2005-2012 JJA	2013-2014 JJA	4.71	12.0
Fresno-Drummond **(a)**	6 x 11	1985-2012 JJA	2013-2014 JJA	4.69	12.9
Fresno-Drummond **(b)**	4 x 6	1995-2012 JJA	2013-2014 JJA	4.38	6.9
Fresno-Drummond **(c)**	7 x 7	2000-2012 JJA	2013-2014 JJA	4.63	11.5
Fresno-Drummond **(d)**	5 x 3	2005-2012 JJA	2013-2014 JJA	5.46	21.8

Fresno-SSP **(a)**	6 x 11	1985-2012 JJA	2013-2014 JJA	4.70	19.8
Fresno-SSP **(b)**	4 x 6	1995-2012 JJA	2013-2014 JJA	4.03	5.6
Fresno-SSP **(c)**	7 x 7	2000-2012 JJA	2013-2014 JJA	4.45	15.8
Fresno-SSP **(d)**	5 x 3	2005-2012 JJA	2013-2014 JJA	4.45	14.7
Hanford **(a)**	4 x 6	1994-2012 JJA*	2013-2014 JJA	4.32	7.7
Hanford **(b)**	7 x 7	1998-2012 JJA*	2013-2014 JJA	4.68	10.6
Hanford **(c)**	9 x 6	1998-2012 JJA*	2013-2014 JJA	4.50	6.3
Hanford **(d)**	3 x 7	1998-2012 JJA*	2013-2014 JJA	4.19	6.6
Hanford **(e)**	5 x 3	2003-2012 JJA*	2013-2014 JJA	4.21	7.7
Visalia – N. Church **(a)**	6 x 11	1985-2012 JJA	2013-2014 JJA	5.34	23.6
Visalia – N. Church **(b)**	4 x 6	1995-2012 JJA	2013-2014 JJA	5.20	28.9
Visalia – N. Church **(c)**	7 x 7	2000-2012 JJA	2013-2014 JJA	6.10	32.3
Visalia – N. Church **(d)**	5 x 3	2005-2012 JJA	2013-2014 JJA	5.14	22.5
Oildale **(a)**	6 x 11	1986-2012 JJA**	2013-2014 JJA	6.30	33.2
Oildale **(b)**	4 x 6	1994-2012 JJA**	2013-2014 JJA	6.33	42.1
Oildale **(c)**	7 x 7	1999-2012 JJA**	2013-2014 JJA	6.10	35.6
Oildale **(d)**	5 x 3	2004-2012 JJA**	2013-2014 JJA	5.09	29.2
Greeley-WTC **(a)**	3 x 7	2002-2012 JJA	2013-2014 JJA	3.61	7.9
Greeley-WTC **(b)**	5 x 3	2005-2012 JJA	2013-2014 JJA	3.71	8.5
Platteville **(a)**	3 x 7	2002-2012 JJA	2014 JA***	3.85	4.7
Platteville **(b)**	5 x 3	2005-2012 JJA	2014 JA***	3.78	4.7

*Here years 2008 and 2009 are missing.

**Here year 2005 is missing.

***Here JA stands for July 13th - August 11th.

To illustrate the testing process with an example, the time series of REGiS probabilistic ozone MDA8 forecasts and the corresponding observations at station FD are presented in Figure 3-2. The prediction is generated from the 1995-2012 training data set (see Table 3-1). Figure 3-2a shows REGiS probabilistic forecasts in terms of percentiles, where grey and yellow colors highlight 5-95 and 25-75 percentiles of the PDFs, respectively. The blue line denotes observations and red dashed line is EPA's exceedance threshold for ozone MDA8. The Figure 3-2b displays the corresponding rank histogram revealing into which percentile regions the observations fall across the PDFs. There are 5 percentile regions that divide the percentiles in the following way: 1-20, 21-40, 41-60, 61-80 and 81-100. The bars of the histogram show how many specific observation points (cases), out of all the days in JJA 2013-2014, fall into the abovementioned five percentile regions. For the station FD, the rank histogram is slightly skewed to the left (the model is biased high), with a RMSD of 6.9 cases. The CRPS, which indicates reliability, resolution and uncertainty of a forecast, is 4.38 ppbv and is on the lower side when compared with the CRPSs from the other stations used in this work.

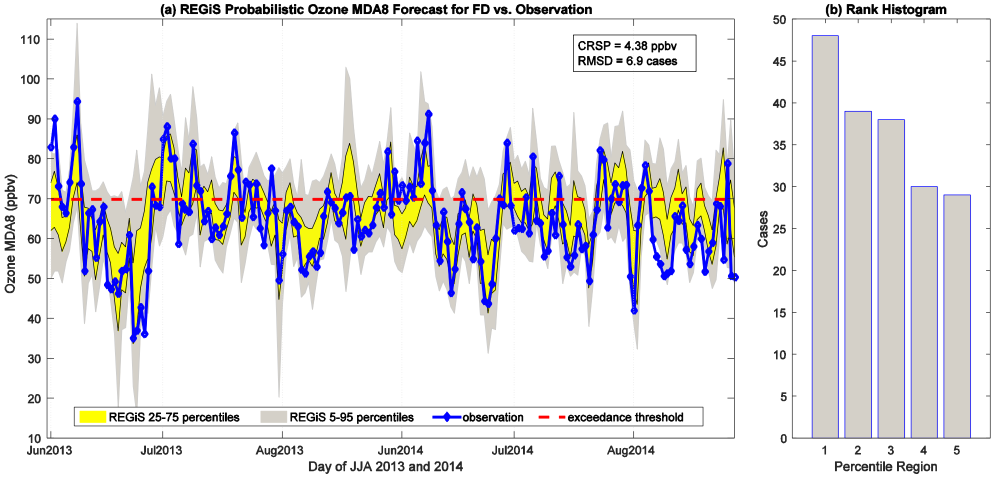


Figure 3-2. (a) Time series of REGiS probabilistic ozone MDA8 forecast for independent data set of JJA 2013-2014 at FD AQ station. (b) A rank histogram corresponding to the probabilistic forecasts in (a).

Further examination of Table 3-1 reveals that the RMSD and CRPS generally increase (skill lowers) for the stations that are farther away from the KFAT and Parlier, sites used to initially tune REGiS parameters $\alpha, \beta, \sigma, \gamma, R_c$ and SOM_{dim}. Specifically, the 2 lowest RMSD values belong to AQ stations FD and Fresno-SSP, where the lowest CRPS occurs at the latter station. These two stations used KFAT as their meteorological reference (see Table 2-2). The RMSD begins to increase as we apply REGiS to other AQ stations, where KVIS and KBFL are used as meteorological references. One potential reason for this RMSD increase is examined below by looking more closely at AQ site Oildale.

Figure 3-3 presents REGiS evaluation at Oildale for the two training periods. Oildale (b) REGiS (see Table 3-1) is developed using 1994-2012 JJA data in contrast to Oildale (d) REGiS that is developed using 2004-2012 JJA data. From Figures 3-3a and 3-3b it is possible to notice that Oildale (b) REGiS forecast is biased high, with observations mainly falling into percentile regions 1 and 2 leading to RMSD = 42.1 cases, the highest RMSD in this study (Figure 3-3b). The likely cause of this bias is the sudden decrease in ozone MDA8 levels at Oildale beginning in 2009, which seem to be noticeably lower than in the years before. Detrending the 1994-2012 MDA8 data does not help in this case because the change in ozone MDA8 with time is not gradual. Figures 3-3c and 3-3d still indicate a significant positive bias in REGiS with RMSD = 29.2 cases; however, the bias is decreased by over 10 cases in comparison with Oildale (b) REGiS. Ozone MDA8 data from 2004-2012 JJA seem to be more representative of the ozone MDA8 observations in 2013-2014 JJA. So, although the REGiS tuning process at Parlier implied that the best training periods are 1995-2012 and 2000-2012, that may not always be the case as the current example with Oildale AQ site illustrates.

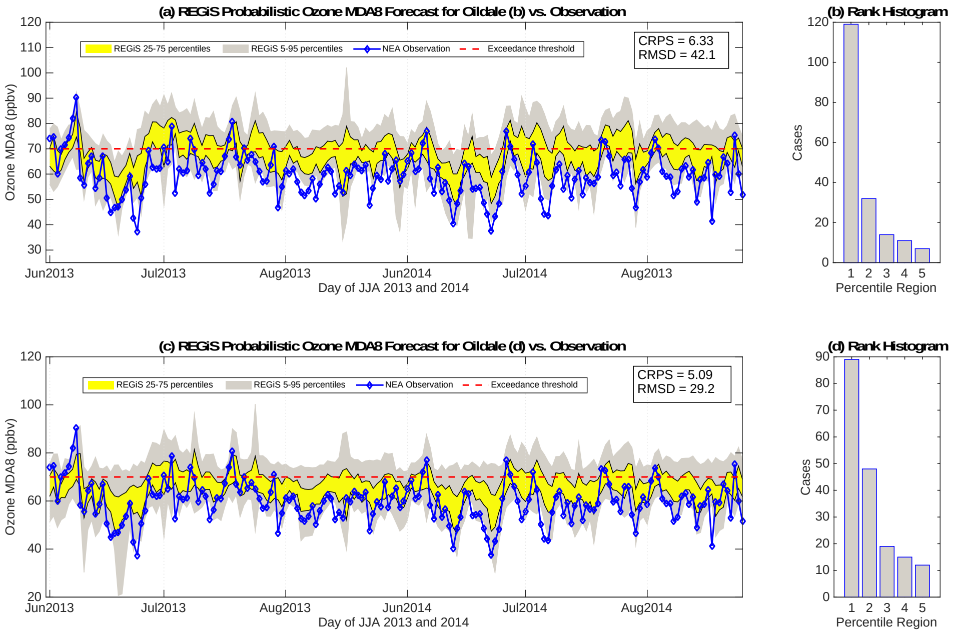


Figure 3-3. (a) Same as in Figure 3-2a but for AQ station Oildale (b) as specified in Table 3-1. **(b)** Same as in Figure 3-2b but for AQ station Oildale (b) as specified in Table 3-1. **(c)** Same as in Figure 3-2a but for AQ station Oildale (d) as specified in Table 3-1. **(d)** Same as in Figure 3-2b but for AQ station Oildale (d) as specified in Table 3-1.

Using the settings determined at Parlier, REGiS is also evaluated for northeastern CO at Greeley-WTC, a long-term AQ station, and a short-term DSICOVER-AQ campaign site – Platteville (Crawford and Pickering 2014). Platteville is used to test whether REGiS ozone MDA8 predictions for an AQ station (in this case Greeley-WTC) are representative of its surroundings. It is located about 30 km south of Greeley-WTC (Figure 2-5) with available ozone data from July 13th to August 11th of 2014. For both stations REGiS is trained using Greeley-Weld County Airport (KGXY) meteorological station and Greeley-WTC AQ site. At Greeley-WTC, RMSD and CRPS values are comparable to those of the sites at SJV (Table 3-1) indicating that REGiS is able to perform in CO at least as well as it does in SJV. Although RMSD values at Platteville indicate a better REGiS skill than at Greeley-WTC, these values are less reliable for the skill comparison because of the smaller testing data sample size available. Therefore, in this case CRPS is used for the evaluation. Similar CRPS at both of the stations suggests that REGiS ozone prediction for Greeley-WTC is applicable to its surroundings.

Examples described above underscore the importance of the initial REGiS tuning process and the length of the training data. Although tuning REGiS settings at a single AQ station (in this case Parlier) works well for some stations (e.g. FD, Hanford and Greeley-WTC), it does not seem to be the best way to operate REGiS. As illustrated by the Oildale case, for the stations with sudden changes in ozone MDA8 REGiS may perform better using smaller and more recent training periods. In real-time forecasting situations, it would be ideal to tune REGiS separately for each AQ station of interest for the best results. That would entail running the sensitivity analysis presented in Chapter 2. Various sets of training data could be tested, but it is recommended to use at least 4 years of data to make sure that the sample size is sufficient for SOMs clustering procedure. It is best to train REGiS on a specific season, such as here where the summer data is used. Similarly, the data could be separated into the other seasons: Autumn, Winter and Spring. The meteorological station that supplies predictors of MDA8 ozone for REGiS should be supported by MOS products to allow REGiS to run operationally.

In the presented REGiS testing experiment presented here, it is assumed that the meteorology is known perfectly and hence ERA-Interim and NCDC observations are used to make MDA8 ozone predictions. In operational forecasting this would not be the case and the meteorological data would have to be supplied by NWP systems such as GFS and MOS in order to run REGiS. This process is demonstrated in Chapter 4; here the feasibility of the REGiS design is demonstrated. The results of this analysis indicate that REGiS could be useful to a forecaster if properly tuned. The RMSD values at the tested sites suggest that generally REGiS is able to estimate reliable distribution of potential MDA8 ozone mixing ratios. By itself, REGiS may not provide an AQ forecaster with a definitive guidance, but together with NAQFC and possibly other tools, it would be easier to make a well-informed decision regarding the predicted AQI category for a specific station.

Chapter 4

Operational REGiS

Because REGiS is meant to be a real-time ozone forecasting model, it is important to evaluate the model in its operational mode. Although the model as described below has been operational only for a few months, it is still possible to glean some useful information from the verification analysis. The analysis presented here is preliminary; nevertheless, it lays out the foundation for the further study that is imperative if REGiS is to become a fully operational model.

4.1 Operational REGiS Configuration

To verify REGiS in an operational environment the model is applied to the Philadelphia area, which is also referred to as Southeast PA. Currently the region contains eight AQ sites that measure hourly ozone. The goal of an AQ forecaster assigned to Philadelphia area is to predict MDA8 ozone for the region as a whole. The station with the highest MDA8 ozone counts as the verifying observation for the region. Every day the forecasters provide a two-day forecast. The reason for choosing Southeast PA as a verification location for REGiS is based on the author's familiarity with the region's AQ forecasters.

In their typical routine on a given day, AQ forecasters produce a forecast for the following two days and submit it to https://www.airnow.gov/. The AQ forecast considers possible exceedances for all of the criteria pollutants and expresses its prediction using AQI metric (see

Chapter 1), where the most common pollutants are PM and ozone. As of now REGiS is only able to predict surface ozone, so the other pollutants are not considered in the presented evaluation.

To test the feasibility of REGiS design, the one-day MDA8 ozone forecasts developed by REGiS were sent to the region's AQ forecasters by the early afternoon on a given day, allowing them to incorporate REGiS' guidance into their final forecast. As mentioned previously, the Southeast PA region contains eight AQ monitoring sites. Based upon personal communication with AQ forecasters, Northeast Philadelphia Airport AQ station (NEA) has been selected as a representative site of the region. The ozone data from NEA and meteorological data from the Philadelphia Northeast Airport weather station (KPNA) are used to configure REGiS. Configuration of this version of REGiS consists of the training and prediction data.

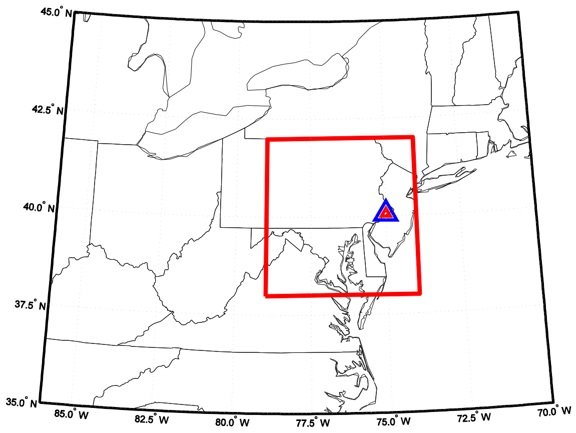
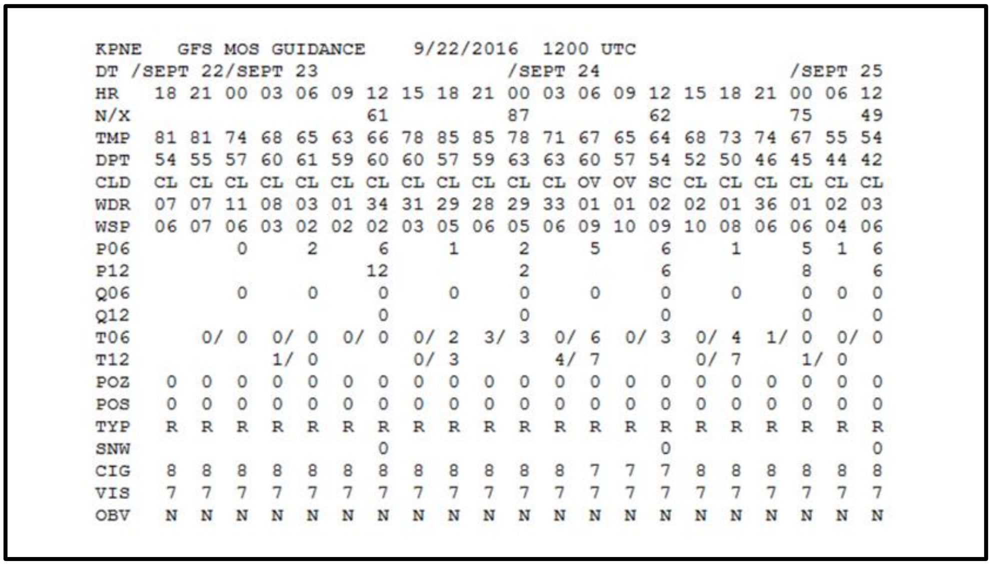


Figure 4-1. Domain of the operational Southeast PA region used by REGiS. The red square outlines the secondary REGiS domain (for more details see Chapter 2). The red triangle indicates location of the meteorological station KPNE and the blue triangle indicates the location of NEA AQ site.

Meteorological and ozone training data come from NCDC and AQS databases respectively. Synoptic training data comes from the 2nd-generation NOAA Global Ensemble Forecast System (GEFS) Reforecast control run, which provides consistency because the prediction synoptic data also comes from the GFS (operational). GFS is a 4-D weather prediction model that is operated by the National Centers for Environmental Prediction (NCEP, Environmental Modeling Center 2003). GEFS Reforecast v2 is a dataset with historical weather forecasts; it is available at http://www.esrl.noaa.gov/psd/forecasts/reforecast2/. The domains used by REGiS with GEFS Reforecast v2 are shown in Figure 4-1. Training data is used to train the model as described in Chapter 2. The training period used for the predictions is May-June 2004-2014 before the July of 2016 and June-July-August 2004-2014 after July of 2016 (see section 4.2 for more details).

Prediction data is used as an input for the REGiS to produce MDA8 ozone forecast (Figure 2-9). Prediction data for the meteorological regression predictors comes from GFS and the National American Mesoscale Forecast System (NAM) MOS (Carter et al. 1989; Rogers et al. 2005). MOS method applies a regression model trained on the archived model data from GFS and NAM to adjust the station-specific meteorological forecasts from operational GFS and NAM. REGiS averages afternoon relevant meteorological variables that are produced by MOS output. An example of the GFS MOS output for KPNE is illustrated in the Figure 4-2. Current MOS data are available at http://www.weather.gov/. Spatial GFS data for pattern determination can be acquired from http://www.emc.ncep.noaa.gov/. Operational REGiS configuration is summarized in the Figure 4-3.



```
KPNE   GFS MOS GUIDANCE    9/22/2016  1200 UTC
DT /SEPT 22/SEPT 23               /SEPT 24                /SEPT 25
HR   18 21 00 03 06 09 12 15 18 21 00 03 06 09 12 15 18 21 00 06 12
N/X                    61          87          62          75    49
TMP  81 81 74 68 65 63 66 78 85 85 78 71 67 65 64 68 73 74 67 55 54
DPT  54 55 57 60 61 59 60 60 57 59 63 63 60 57 54 52 50 46 45 44 42
CLD  CL CL CL CL CL CL CL CL CL CL CL CL OV OV SC CL CL CL CL CL CL
WDR  07 07 11 08 03 01 34 31 29 28 29 33 01 01 02 02 01 36 01 02 03
WSP  06 07 06 03 02 02 02 03 05 06 05 06 09 10 09 10 08 06 06 04 06
P06         0     2     6     1     2     5     6     1     5  1  6
P12                    12           2           6           8     6
Q06         0     0     0     0     0     0     0     0     0  0  0
Q12                     0           0           0           0     0
T06      0/ 0  0/ 0  0/ 0  0/ 2  3/ 3  0/ 6  0/ 3  0/ 4  1/ 0  0/ 0
T12               1/ 0        0/ 3        4/ 7        0/ 7     1/ 0
POZ   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
POS   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
TYP   R  R  R  R  R  R  R  R  R  R  R  R  R  R  R  R  R  R  R  R  R
SNW                     0                       0                 0
CIG   8  8  8  8  8  8  8  8  8  8  8  8  7  7  7  8  8  8  8  8  8
VIS   7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7
OBV   N  N  N  N  N  N  N  N  N  N  N  N  N  N  N  N  N  N  N  N  N
```

Figure 4-2. An example of the GFS MOS output for KPNE.

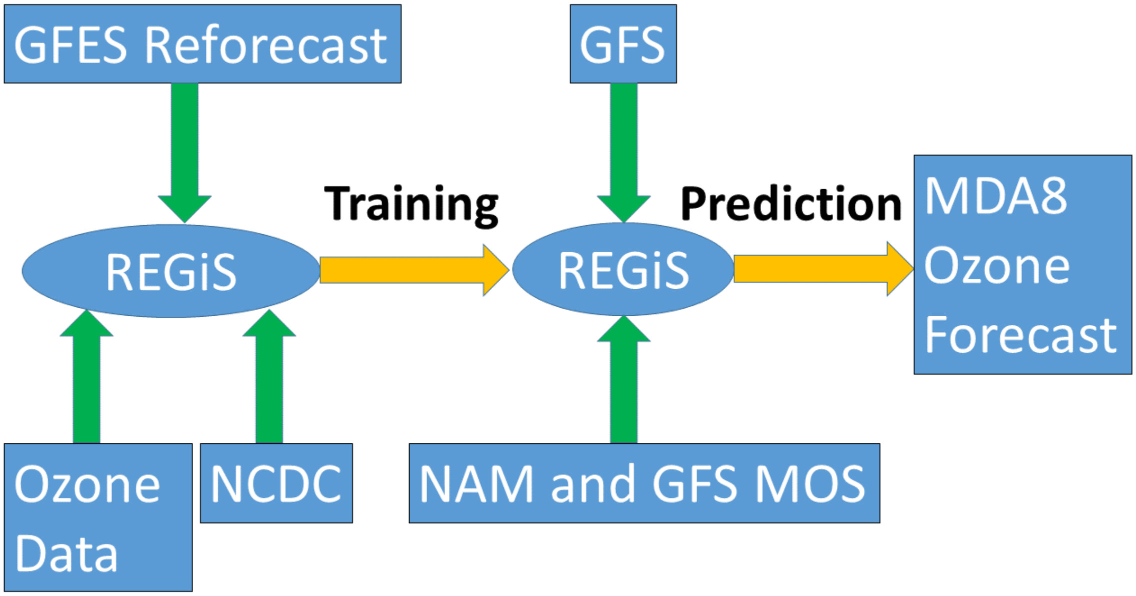


Figure 4-3. Diagram summarizing the process that REGiS performs in order to generate a probabilistic MDA8 ozone forecast.

4.2 REGiS Philadelphia Experiment

For the following experiment, the REGiS domain is shown in Figure 4-1 where the AQ monitoring station is NEA and the meteorological station is KPNE. The model has been operated in two phases: for May-June of 2016 and for July-August-September of 2016. Although in principle the model has been operational over the five months, many of the forecasts are missing. The main reason for the missing forecasts is the lack of the automation in the model. Although the model only takes a few minutes to produce a forecast, the process has to be launched by an operator. Additionally, in its initial stages the model experienced a number of crashes for various reasons that had to be addressed. The constant modifications to the model and the limited availability of an operator contributes significantly to the reduced number of the available REGiS forecasts. For the verification of REGiS only the data from phase one is used consisting of May-June from 2016. The second phase of REGiS operation has been marred with too many errors to qualify for proper evaluation. Nevertheless, the data is briefly analyzed in the next section helping to establish the goals for future studies.

For the experiment with the phase one data, the model is not tuned as has been done for the CA REGiS evaluation (see Chapter 2). Ideally the REGiS needs to be tuned before it is used, but this has not been possible in the current situation due to the logistical reasons. Therefore, various settings are tried throughout the experiment. For the planned future applications, the model will be tuned using the method explained in Chapter 2. Because of the conditions under which the REGiS application is tested, the results are preliminary and do not serve as a proper quantification of the REGiS performance.

The results of the experiment are shown in the Figure 4-4. Although considerable data is missing, the model is able to capture the general ozone variability with a remarkably low RMSD score of 3.6 cases. CRPS is on the high side in comparison with the results in Table 3-1. The likely reason for the high CRPS in the given situation is the small sample size of the data used for verification. CRPS can be separated into the three terms known as reliability, resolution and uncertainty (Hersbach 2000). Low RMSD implies that the reliability is not the main issue in CRPS meaning that the resolution and uncertainty are the components of CRPS that are keeping it high (Wilks 2011). Low number of exceedances in the evaluation data supports this line of thinking: there are not enough observations of exceedances and the corresponding REGiS forecasts to further “resolve” our knowledge regarding the model skill. More data is needed to make any additional conclusions regarding the REGiS performance, but the available data implies that REGiS potentially could be useful (i.e. compete with skill against other methods).

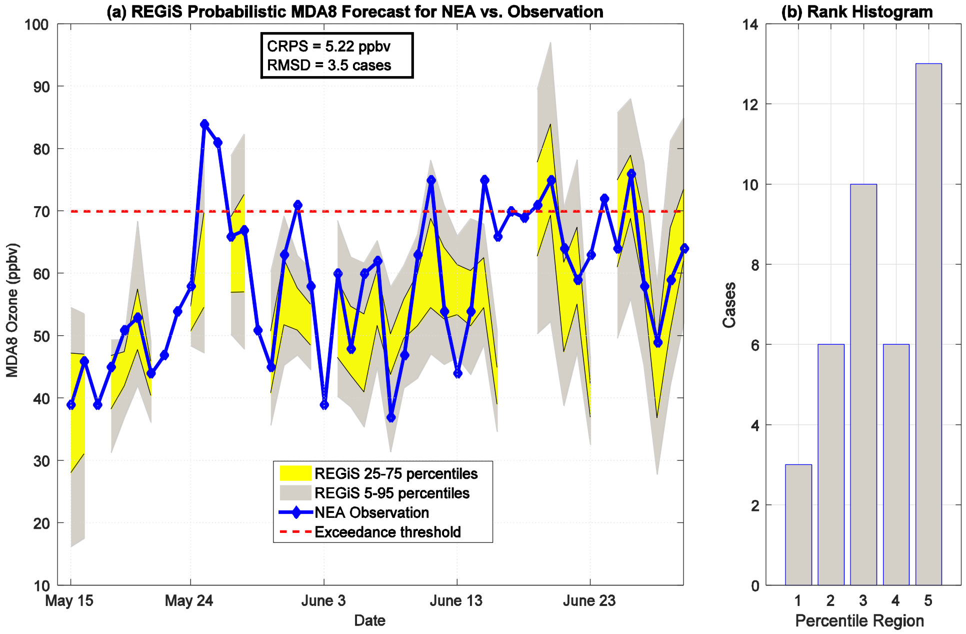


Figure 4-4. (a) Same as in Figure 3-2a but for AQ station NEA. The dates refer to year 2016. **(b)** Same as in Figure 3-2b but for AQ station NEA.

4.3 Future Directions

The main goal of this thesis has been the assessment of the REGiS feasibility. The natural continuation trajectory of this work, however, would be to compare REGiS with the other ozone forecasting methods in order to determine if the REGiS forecasts are useful operationally. Although it is not possible to carry out such complete analysis with the given data, it is possible to gain an idea of how to go about making such analysis.

In the previous section, it has been mentioned that the REGiS completed forecasts for July-August-September of 2016. Initially these forecasting data were not intended for the analysis, so the available predictions are intermittent. For about the third of these forecasts, NAM MOS guidance was not available on NOAA File Transfer Protocol (FTP) server and thus has been omitted from the list of the REGiS inputs. Nevertheless, the data can help to increase the sample size of the REGiS probabilistic predictions and give more information regarding the REGiS performance.

As discussed in Chapter 1, an ozone exceedance occurs when MDA8 ozone crosses the threshold of 70 ppbv. Using REGiS probabilistic forecasts (PDFs) it is possible to estimate the probability of this exceedance given the forecast. This process is completed for all of the available REGiS forecasts using the data (N = 73) from the two phases of the REGiS Philadelphia experiment (May through September of 2016). The results are visualized using the attributes diagram, which is shown in the Figure 4-5. The attributes diagram captures the full joint distribution of the probabilistic forecasts for a binary event (in this case exceedance or non-exceedance) and the conditional distribution of the observations given the corresponding probabilistic forecast (Hsu and Murphy, 1986). The 1:1 line in the figure indicates forecasts with the perfect reliability – the forecasted frequency of an event is matched by the observed frequency. The line labeled “no skill” indicates that the resolution of the forecast is smaller than the reliability of the forecast as defined by the Brier Skill Score (for more details see Chapter 8 in Wilks, 2011). "No resolution" horizontal line marks the region where the forecast is unable to give more information than the climatology. From the figure it is possible to observe some serious problems with the REGiS forecasts. In the dataset used, there are only a few cases available where the probabilistic forecasts indicate the chance of an exceedance to be above 70% and none of these forecasts verify. There are more indications of the overconfidence in the forecasts in the 30-50% range. Although this analysis is only exploratory, as mentioned previously, there is enough evidence here to see that the REGiS has a tendency of over-prediction. One reason for this is the unbalanced proportion of non-exceedances to exceedances in the dataset. It is possible that with the larger sample size the result would improve.

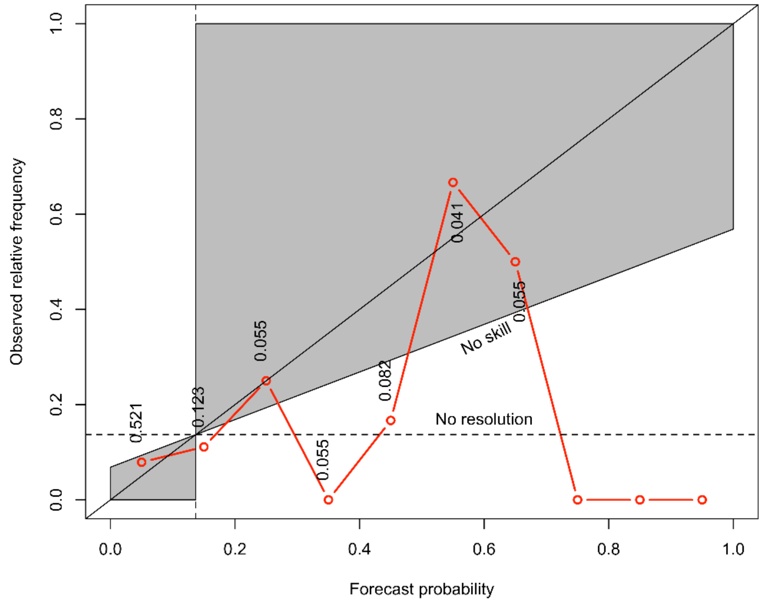


Figure 4-5. Attributes diagram for the REGiS probabilistic forecasts carried out for Southeast PA region during the selected days from May-September of 2016. The values shown above or below the red circles indicate the frequency of the data available for that particular evaluation. The three unlabeled red points have the values of 0.014, 0.041 and 0.014 respectively. For more details see text.

It is possible to compare REGiS with the NAQFC (for more details see Chapter 1). The data for the NAQFC (the same dates as used by the REGiS above) is supplied by the Southeast PA AQ forecasters. The discrete NAQFC MDA8 ozone forecasts are converted into binary forecasts where, NAQFC > 71 ppbv forecasts are counted as probability = 1 and NAQFC ≤ 71 ppbv forecasts are counted as probability = 0 (based on the methodology by Garner and Thompson 2013). The results are illustrated in the Figure 4-6 and indicate that the NAQFC model is unable to skillfully predict exceedances. The forecast of an exceedance verifies only about 30% of the time.

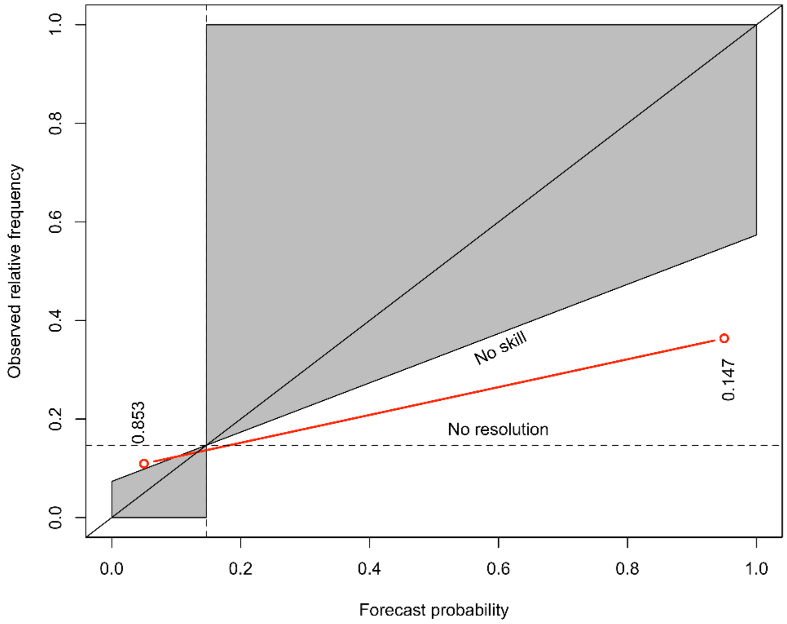


Figure 4-6. Attributes diagram as in the Figure 4-5 but for the NAQFC modeling system.

Because the comparison between the models performed above is generated using only a limited dataset, it is difficult to make any strong conclusions regarding the results. Nonetheless, such analysis indicates that the more structured study in this direction can be useful in comparing the skill level of the REGiS and NAQFC. There is a suggestion that both of the models struggle with capturing the exceedances effectively; however, using only the attributes diagram to diagnose the models of interest may be insufficient and the other methods would have to be explored (Wilks 2011).

Another possible direction for comparing REGiS with other ozone prediction approaches lies in transforming REGiS forecasts into discrete values. Figure 4-7 shows the evaluation of the four different MDA8 ozone forecasting methods for the selected data from May-September of 2016. Discrete REGiS forecasts are the combination of the ozone MDA8 prediction PDFs averages and the multivariate linear regression equation that uses REGiS training data to produce its forecasts. NOAA model refers to NAQFC. Human forecasts are the integrated forecasts generated by the Southeast PA region AQ forecasters. Statistical guidance is the multivariate linear regression model developed by the Southeast PA region AQ forecasters.

The comparison of the models demonstrates that the human forecasters are noticeably more skillful than the other three approaches with the mean square error (MSE, Wilks 2011) being around 64, while the other methods are in the 103-105 range. Although the discrete REGiS differs from the probabilistic REGiS, it is possible to note that it produces false alarms (overforecasting) more than any of the other forecasting approaches. This is consistent with what has been observed in the attributes diagram in the Figure 4-5 for the probabilistic REGiS, which is also overconfident. On the positive side, MSEs for the discrete REGiS, NAQFC and the statistical guidance are comparable meaning that the current REGiS is not necessarily worse than the current operational methods.

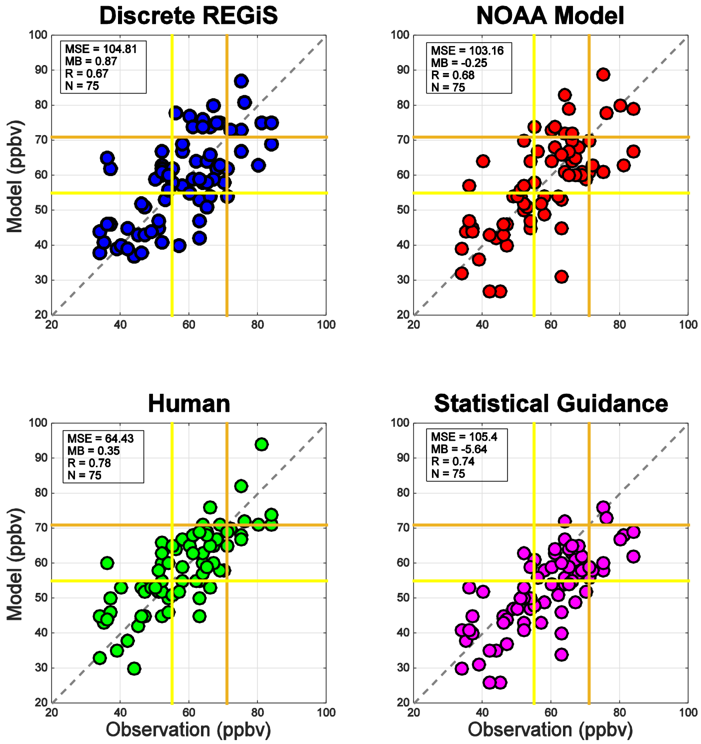


Figure 4-7. Evaluation of the MDA8 ozone forecasts for the selected days in May-September of 2016 generated by discrete REGiS, NAQFC (or NOAA model), humans and statistical guidance (developed by Philadelphia region AQ forecasters).

To further investigate the question of the REGiS verification, another separate study is required. In that study the REGiS needs to be properly tuned as described in Chapter 2 before it is subjected to an evaluation. Improvements to the REGiS methodology such as domain and predictor variable choice should also be explored. In the current version of the REGiS the uncertainty of the meteorological inputs has not been addressed, even though the REGiS relies on the accurate meteorology to make a skillful prediction. It is recommended that the inputted meteorological data is perturbed in some way to account for the intrinsic uncertainty in the atmospheric processes.

\

Chapter 5

Conclusion

In this thesis, a novel statistical approach is developed to produce probabilistic daily surface ozone MDA8 forecasts. The model draws an inspiration from the tree-based and stratified models that exploit the fact that the association between an air pollutant and meteorology may be different in different meteorological regimes. Meteorological regimes are identified by SOM, an ANN technique for pattern recognition. Once regimes are identified, a stepwise weighted quadratic regression equation is developed for each weather pattern. The SOM method allows both identification of different meteorological regimes and the grouping of them according to similarity. In this way, when a regression is developed for a specific regime, data from all the other relevant regimes can also be used, with weights based on the similarity between the regimes. This approach yields a distinct model for each regime while still taking into account all relevant training cases when building each regime's model. All of the resultant regression models are combined together to produce a PDF of a MDA8 ozone forecast using kernel density smoothing. The model is named REGiS and derives its name from the three words: regression in SOM.

REGiS is evaluated at SJV, CA, a location known for its poor AQ in the US, and northeastern CO to demonstrate diverse applicability of the model. Before REGiS can be evaluated it needs to be tuned. REGiS is tuned at the Parlier AQ station using the meteorological data from the nearby meteorological station KFAT. Four training periods are used to study REGiS sensitivity to the change in emissions. Once tuned, REGiS is tested at the several independent AQ stations in order to verify its ability to produce reliable probabilistic forecasts. Two skill metrics are used to evaluate REGiS: CRPS and RMSD from flat rank histogram. CRPS determines reliability, resolution and uncertainty of a forecast, while RMSD determines whether a predicted PDF is representative of the true distribution. Results indicate that for AQ stations located near Parlier, which used KFAT to develop regressions, REGiS achieves lower RMSD values than for the stations that are further away and which used other meteorological sites for the regressions. CRPS behaves similarly with the exception of Hanford AQ station. This implies that tuning REGiS for each station separately would generate a more robust model for the probabilistic MDA8 ozone prediction.

Additionally, the real-time forecasting experiment is performed in the Southeast PA region (Philadelphia area) using the NEA AQ monitor next to the KPNE meteorological station. The results are promising, with a low RMSD value indicating that the REGiS is able to capture some of the MDA8 ozone variability due to meteorology. There is some preliminary evidence indicating that the REGiS tends to over-predict ozone MDA8 leading to more “false alarm” forecasts than other ozone MDA8 prediction approaches. However, further study is required to properly evaluate REGiS.

The uniqueness of the REGiS is its ability to generate probabilistic MDA8 ozone forecasts. Uncertainty quantification using a PDF gives an advantage to a probabilistic forecast over discrete deterministic forecast. The PDF allows an AQ forecaster to see whether a given meteorological setting is favorable or not for an ozone pollution episode. REGiS is not designed to account for sudden local emission changes or events such as biomass fires, but by using its large historical database REGiS is well suited for informing the probability of an AQI category given a particular meteorological set up. This thesis underlines the value of the REGiS ozone MDA8 probabilistic forecasts and their ability to aid AQ forecasters by quantifying the uncertainty of the real-time ozone prediction.

References

Abdulkader, H., and D. Roviras, 2012: Generating cryptography keys using self-organizing maps. *2012 International Symposium on Wireless Communication Systems (ISWCS)*, 736-740.

Agarwal, P., and A. Skupin, 2008: *Self-organising maps: Applications in geographic information science.* John Wiley & Sons.

Aw, J., and M. J. Kleeman, 2003: Evaluating the first-order effect of intraannual temperature variability on urban air pollution. *Journal of Geophysical Research: Atmospheres*, **108**(D12), 4365, doi:10.1029/2002JD002688.

Beaver, S., and A. Palazoglu, 2009: Influence of synoptic and mesoscale meteorology on ozone pollution potential for San Joaquin Valley of California. *Atmospheric Environment*, **43,** 1779-1788.

Bowman, A. W., and A. Azzalini, 1997: *Applied smoothing techniques for data analysis: The kernel approach with S-Plus illustrations: The kernel approach with S-Plus illustrations.* OUP Oxford.

Brook, R. D., J. R. Brook, B. Urch, R. Vincent, S. Rajagopalan, and F. Silverman, 2002: Inhalation of fine particulate air pollution and ozone causes acute arterial vasoconstriction in healthy adults. *Circulation*, **105,** 1534-1536.

Burrows, W. R., M. Benjamin, S. Beauchamp, E. R. Lord, D. McCollor, and B. Thomson, 1995: CART decision-tree statistical analysis and prediction of summer season maximum surface ozone for the Vancouver, Montreal, and Atlantic regions of Canada. *Journal of applied meteorology*, **34,** 1848-1862, doi: http://dx.doi.org/10.1175/1520-0450(1995)034<1848:CDTSAA>2.0.CO;2.

Carlson, T. N., 1991: *Mid-latitude Weather Systems*. Harper Collins Academic.

Carmichael, G. R., A. Sandu, T. Chai, D. N. Daescu, E. M. Constantinescu, and Y. Tang, 2008: Predicting air quality: Improvements through advanced methods to integrate models and measurements. *Journal of Computational Physics*, **227,** 3540-3571, doi:10.1016/j.jcp.2007.02.024.

Carter, G. M., J. P. Dallavalle, and H. R. Glahn, 1989: Statistical forecasts based on the national meteorological center's numerical weather prediction system. *Weather and Forecasting*, **4**, 401-412, http://dx.doi.org/10.1175/1520-0434(1989)004<0401:SFBOTN>2.0.CO;2.

Chai, T., and Coauthors, 2013: Evaluation of the United States National Air Quality Forecast Capability experimental real-time predictions in 2010 using Air Quality System ozone and NO_2 measurements. *Geoscientific Model Development*, **6**, 1831-1850, doi:10.5194/gmd-6-1831-2013.

Cobourn, W. G., 2007: Accuracy and reliability of an automated air quality forecast system for ozone in seven Kentucky metropolitan areas. *Atmospheric Environment*, **41**, 5863-5875, doi:10.1016/j.atmosenv.2007.03.024.

Comrie, A. C., 1997: Comparing neural networks and regression models for ozone forecasting. *Journal of the Air & Waste Management Association*, **47**, 653-663, doi:10.1080/10473289.1997.10463925.

Cooper, O. R., A. O. Langford, D. D. Parrish, and D. W. Fahey, 2015: Challenges of a lowered U.S. ozone standard. *Science,* **348,** 1096-1097, doi:10.1126/science.aaa5748.

Cox, W. M., and S.-H. Chu, 1993: Meteorologically adjusted ozone trends in urban areas: A probabilistic approach. *Atmos. Environ.*,**27,** 425–434, doi:10.1016/0957-1272(93)90019-3.

Crawford, J. H., and K. E. Pickering, 2014: DISCOVER-AQ: Advancing strategies for air quality observations in the next decade. *Environ. Manage*, 4-7.

Cressman, G. P., 1959: An operational objective analysis system. *Monthly Weather Review*, **87,** 367-374. doi: http://dx.doi.org/10.1175/1520-0493(1959)087<0367:AOOAS>2.0.CO;2.

Dabberdt, W. F., and E. Miller, 2000: Uncertainty, ensembles and air quality dispersion modeling: applications and challenges. *Atmospheric Environment*, **34,** 4667-4673.

Dabberdt, W. F., and Coauthors, 2004: Meteorological Research Needs for Improved Air Quality Forecasting: Report of the 11th Prospectus Development Team of the U.S. Weather Research Program*. *Bulletin of the American Meteorological Society*, **85,** 563-586, doi: http://dx.doi.org/10.1175/BAMS-85-4-563.

Dee, D., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, **137,** 553-597, doi:10.1002/qj.828.

Delle Monache, L., and R. B. Stull, 2003: An ensemble air-quality forecast over western Europe during an ozone episode. *Atmospheric Environment,* **37,** 3469-3474.

Delle Monache, L., X. Deng, Y. Zhou, and R. Stull, 2006a: Ozone ensemble forecasts: 1. A new ensemble design. Journal of Geophysical Research: Atmospheres, **111**, D05307, doi:10.1029/2005JD006310.

Delle Monache, L., J. P. Hacker, Y. Zhou, X. Deng, and R. B. Stull, 2006b: Probabilistic aspects of meteorological and ozone regional ensemble forecasts. Journal of Geophysical Research: Atmospheres, **111**, D24307, doi:10.1029/2005JD006917.

Diaz-Robles, L. A., J. C. Ortega, J. S. Fu, G. D. Reed, J. C. Chow, J. G. Watson, and J. A. Moncada-Herrera, 2008: A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: the case of Temuco, Chile. *Atmospheric Environment*, **42,** 8331-8340, doi:10.1016/j.atmosenv.2008.07.020.

Djalalova, I., L. Delle Monache, and J. Wilczak, 2015: PM2.5 analog forecast and Kalman filter post-processing for the Community Multiscale Air Quality (CMAQ) model. *Atmospheric Environment*, **108,** 76-87, doi:10.1016/j.atmosenv.2015.02.021.

Dutot, A.-L., J. Rynkiewicz, F. E. Steiner, and J. Rude, 2007: A 24-h forecast of ozone peaks and exceedance levels using neural classifiers and weather predictions. *Environmental Modelling & Software*, **22,** 1261-1269, doi:10.1016/j.envsoft.2006.08.002.

Environmental Modeling Center, 2003: The GFS atmospheric model. NCEP Office Note 442, Global Climate and Weather Modeling Branch, EMC, Camp Springs, Maryland.

EPA, U. S., 2003: Guidelines for developing an air quality (Ozone and $PM_{2.5}$) forecasting program, 126 pp.

——, 2015: National Ambient Air Quality Standards (NAAQS). [Available online at http://www3.epa.gov/ttn/naaqs/criteria.html.]

——, 2016: Air Quality Index (AQI) basics. [Available online at https://airnow.gov/index.cfm?action=aqibasics.aqi.]

Fenger, J., 2009: Air pollution in the last 50 years – From local to global. *Atmospheric Environment*, **43,** 13-22.

Flynn, J., and Coauthors, 2010: Impact of clouds and aerosols on ozone production in Southeast Texas. *Atmospheric Environment*, **44,** 4126-4133.

Frost, G. J., and J. F. Meagher, 2010: Addressing Scientific Challenges for Air Quality Forecasting: International Workshop on Air Quality Forecasting Research; Boulder, Colorado, 2–3 December 2009. *Eos, Transactions American Geophysical Union*, **91,** 145-145, doi:10.1029/2010EO160008.

Gardner, M. W., and S. R. Dorling, 1998: Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment*, **32,** 2627-2636.

Garner, G. G., and A. M. Thompson, 2012: The Value of Air Quality Forecasting in the Mid-Atlantic Region. *Weather, Climate, and Society*, **4,** 69-79, doi: http://dx.doi.org/10.1175/WCAS-D-10-05010.1.

——, 2013: Ensemble statistical post-processing of the National Air Quality Forecast Capability: Enhancing ozone forecasts in Baltimore, Maryland. *Atmospheric Environment*, **81,** 517-522, doi:10.1016/j.atmosenv.2013.09.020.

Greenbaum, D. S., J. D. Bachmann, D. Krewski, J. M. Samet, R. White, and R. E. Wyzga, 2001: Particulate air pollution standards and morbidity and mortality: case study. *American journal of epidemiology*, **154,** 78-90.

Greybush, S. J., S. E. Haupt, and G. S. Young, 2008: The regime dependence of optimally weighted ensemble model consensus forecasts of surface temperature. *Weather and Forecasting*, **23**, 1146-1161.

Hamill, T. M., 2001: Interpretation of Rank Histograms for Verifying Ensemble Forecasts. *Monthly Weather Review*, **129,** 550-560, doi:http://dx.doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2.

Hersbach, H., 2000: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. *Weather and Forecasting*, **15,** 559-570, doi:http://dx.doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Hewitson, B., and R. Crane, 2002: Self-organizing maps: applications to synoptic climatology. *Climate Research*, **22,** 13-26.

Hewitson, B. C., and R. G. Crane, 2006: Consensus between GCM climate change projections with empirical downscaling: precipitation downscaling over South Africa. *International Journal of Climatology*, **26,** 1315-1337.

Hill, T., L. Marquez, M. O'Connor, and W. Remus, 1994: Artificial neural network models for forecasting and decision making. *International Journal of Forecasting*, **10,** 5-15.

Hsu, W., and A. H. Murphy, 1986: The attributes diagram: A geometrical framework for assessing the quality of probability forecasts. *Int. J. Forecasting*,**2,** 285–293.

Hudson, D. A., and B. C. Hewitson, 2001: The atmospheric response to a reduction in summer Antarctic sea-ice extent. *Climate Research*, **16,** 79-99.

Jenkin, M. E., and K. C. Clemitshaw, 2000: Ozone and other secondary photochemical pollutants: chemical processes governing their formation in the planetary boundary layer. *Atmospheric Environment*, **34,** 2499-2527.

Jensen, A. A., A. M. Thompson, and F. J. Schmidlin, 2012: Classification of Ascension Island and Natal ozonesondes using self-organizing maps. Journal of Geophysical Research: Atmospheres, **117**, D04302, doi:10.1002/2011JD016573.

Jhun, I., B. Coull, A. Zanobetti, and P. Koutrakis, 2015: The impact of nitrogen oxides concentration decreases on ozone trends in the USA. *Air Qual Atmos Health*, **8,** 283-292.

Johnson, N. C., S. B. Feldstein, and B. Tremblay, 2008: The continuum of Northern Hemisphere teleconnection patterns and a description of the NAO shift with the use of self-organizing maps. *Journal of Climate*, **21,** 6354-6371.

Kalnay, E., 2003: *Atmospheric modeling, data assimilation and predictability.* Cambridge university press.

Keith, R., and S. M. Leyton, 2007: An Experiment to Measure the Value of Statistical Probability Forecasts for Airports. *Weather and Forecasting*, **22,** 928-935.

Klimont, Z., and Coauthors, 2009: Projections of SO_2, NO_x and carbonaceous aerosols emissions in Asia. *Tellus B*, **61,** 602-617.

Klonecki, A., and H. Levy, 1997: Tropospheric chemical ozone tendencies in CO-CH_4-NO_y-H_2O system: Their sensitivity to variations in environmental parameters and their application to a global chemistry transport model study. *Journal of Geophysical Research: Atmospheres*, **102,** 21221-21237.

Kohonen, T., 2001: Self-organizing maps, vol. 30 of Springer Series in Information Sciences. Springer Berlin.

——, 2013: Essentials of the self-organizing map. *Neural Networks*, **37,** 52-65.

Lamarque, J. F., and Coauthors, 2010: Historical (1850–2000) gridded anthropogenic and biomass burning emissions of reactive gases and aerosols: methodology and application. *Atmos. Chem. Phys.*, **10,** 7017-7039.

Lee, D. S., M. R. Holland, and N. Falla, 1996: The potential impact of ozone on materials in the UK. *Atmospheric Environment*, **30,** 1053-1065.

Lelieveld, J., and P. J. Crutzen, 1990: Influences of cloud photochemical processes on tropospheric ozone. *Nature*, **343,** 227-233.

Lindblom, B., P. Macneilage, and M. Studdert-Kennedy, 1983: Self-organizing processes and the explanation of phonological universals. *Linguistics*, **181**.

Liu, Y., R. H. Weisberg, and C. N. K. Mooers, 2006: Performance evaluation of the self-organizing map for feature extraction. *Journal of Geophysical Research: Oceans*, **111**, C05018, doi:10.1029/2005JC003117.

McKeen, S., and Coauthors, 2005: Assessment of an ensemble of seven real-time ozone forecasts over eastern North America during the summer of 2004. *Journal of Geophysical Research: Atmospheres*, **110,** D21307, doi:10.1029/2005JD005858.

Murazaki, K., and P. Hess, 2006: How does climate change contribute to surface ozone change over the United States? *Journal of Geophysical Research: Atmospheres*, **111,** D05301, doi:10.1029/2005JD005873.

Pearce, J. L., J. Beringer, N. Nicholls, R. J. Hyndman, P. Uotila, and N. J. Tapper, 2011: Investigating the influence of synoptic-scale meteorology on air quality using self-organizing maps and generalized additive modelling. *Atmospheric Environment*, **45,** 128-136.

Prybutok, V. R., J. Yi, and D. Mitchell, 2000: Comparison of neural network models with ARIMA and regression models for prediction of Houston's daily maximum ozone concentrations. *European Journal of Operational Research*, **122,** 31-40.

Richardson, A., C. Risien, and F. Shillington, 2003: Using self-organizing maps to identify patterns in satellite imagery. *Progress in Oceanography*, **59,** 223-239.

Richkind, K. E., and A. D. Hacker, 1980: Responses of natural wildlife populations to air pollution. *Journal of Toxicology and Environmental Health*, **6,** 1-10.

Rogers, E., and Coauthors, 2005: The NCEP North American mesoscale modeling system: Final Eta Model/analysis changes and preliminary experiments using the WRF-NMM. Preprints, 21st Conf. on Weather Analysis and Forecasting/17th Conf. on Numerical Weather Prediction, Washington, DC, Amer. Meteor. Soc., 4B.5. [Available online at http://ams.confex.com/ams/WAFNWP34BC/techprogram/paper_94707.htm.]

Ryan, W. F., 1995: Forecasting severe ozone episodes in the Baltimore metropolitan area. *Atmospheric Environment*, **29,** 2387-2398.

Seinfeld, J. H., and S. N. Pandis, 2012: *Atmospheric chemistry and physics: from air pollution to climate change.* John Wiley & Sons.

Shad, R., M. S. Mesgari, and A. Shad, 2009: Predicting air pollution using fuzzy genetic linear membership kriging in GIS. *Computers, Environment and Urban Systems*, **33,** 472-481.

Shahgedanova, M., and T. P. Burt, 1994: New data on air pollution in the former Soviet Union. *Global Environmental Change*, **4,** 201-227.

Sillman, S., and P. J. Samson, 1995: Impact of temperature on oxidant photochemistry in urban, polluted rural and remote environments. *Journal of Geophysical Research: Atmospheres*, **100,** 11497-11508.

Stajner, I., and Coauthors, 2014: National Air Quality Forecast Capability: Status and Research Needs. *AGU Fall Meeting Abstracts*, 3254.

Stauffer, R. M., A. M. Thompson, and G. S. Young, 2016: Tropospheric ozonesonde profiles at long-term U.S. monitoring sites: 1. A climatology based on self-organizing maps. *Journal of Geophysical Research: Atmospheres***, 121**, 1320–1339, doi:10.1002/2015JD023641.

Thompson, A. M., 1984: The effect of clouds on photolysis rates and ozone formation in the unpolluted troposphere. *Journal of Geophysical Research: Atmospheres*, **89,** 1341-1349.

Thompson, M., J. Reynolds, L. H. Cox, P. Guttorp, and P. D. Sampson, 2001: A review of statistical methods for the meteorological adjustment of tropospheric ozone. *Atmospheric Environment*, **35,** 617-630.

Tu, J., Z.-G. Xia, H. Wang, and W. Li, 2007: Temporal variations in surface ozone and its precursors and meteorological effects at an urban site in China. *Atmospheric Research*, **85,** 310-337.

Van der Wal, J., and L. Janssen, 2000: Analysis of spatial and temporal variations of PM 10 concentrations in the Netherlands using Kalman filtering. *Atmospheric Environment*, **34,** 3675-3687.

Vautard, R., and Coauthors, 2009: Skill and uncertainty of a regional air quality model ensemble. *Atmospheric Environment*, **43,** 4822-4832.

Vesanto, J., J. Himberg, E. Alhoniemi, and J. Parhankangas, 2000: *SOM toolbox for Matlab 5.* Citeseer.

Wilks, D. S., 2011: *Statistical methods in the atmospheric sciences.* Vol. 100, Academic press.

Wilczak, J. M., and Coauthors, 2009: Analysis of regional meteorology and surface ozone during the TexAQS II field program and an evaluation of the NMM-CMAQ and WRF-Chem air quality models. *Journal of Geophysical Research: Atmospheres*, **114**, D00F14, doi:10.1029/2008JD011675.

Wilson, L. J., S. Beauregard, A. E. Raftery, and R. Verret, 2007: Calibrated Surface Temperature Forecasts from the Canadian Ensemble Prediction System Using Bayesian Model Averaging. *Monthly Weather Review*, **135,** 1364-1385.

Zhang, Y., M. Bocquet, V. Mallet, C. Seigneur, and A. Baklanov, 2012a: Real-time air quality forecasting, part I: History, techniques, and current status. *Atmospheric Environment*, **60,** 632-655.

——, 2012b: Real-time air quality forecasting, part II: State of the science, current research needs, and future prospects. *Atmospheric Environment*, **60,** 656-676.

VITA

Nikolay Balashov

Education:

The Pennsylvania State University; BS in Meteorology; December 2012
The Pennsylvania State University; BM in Music Composition; December 2012
The Pennsylvania State University; MS in Meteorology; December 2012
The Pennsylvania State University; PhD in Meteorology; December 2016

Professional Experience:

Teaching Assistant, Pennsylvania State University (2015-2016)
Graded assignments and held office hours for *Principles of Atmospheric Measurements* class; helped teach and assisted in planning computer projects for *Application of Computers to Meteorology* class

Research Assistant, Pennsylvania State University (2011-present)
Help with various data analysis and research tasks such as data reading algorithms, statistical analyses, graphics for presentations and publications, writing and proofreading, literature search, and mentoring

Participant, NASA DISCOVER-AQ Field Research Deployment, Platteville, Colorado (Summer 2014)
Assisted with preparing and lunching ozonesondes, assisted in collection of air quality data from trace gas instruments, with calibrations, and data archiving

Participant, NASA SEAC4RS campaign, Houston, Texas (Summer 2013)
Assisted with preparing and lunching ozonesondes and cryogenic frost point hygrometer sondes, analyzed measured data

Intern, The New York Times, The Weather Page (2008-2009)
Weather forecasting, graphics, and short discussions for The New York Times newspaper

Publications:

Balashov, N. V., A. M. Thompson, *and* G. S. Young *(2016),* Probabilistic forecasting of surface ozone with a novel statistical approach, J. Appl. Meteor. Climo., JAMC-D-16-0110, *submitted*.

Balashov, N. V., A. M. Thompson, S. J. Piketh, *and* K. E. Langerman *(2014),* Surface ozone variability and trends over the South African Highveld from 1990 to 2007, *J. Geophys. Res. Atmos.*, 119, 4323–4342*, doi:10.1002/2013JD020555.*

Thompson, A. M., **N. V. Balashov**, J. C. Witte, J. G. R. Coetzee, V. Thouret, and F. Posny (2014), Tropospheric ozone increases over the southern Africa region: bellwether for rapid growth in Southern Hemisphere pollution?, *Atmos. Chem. Phys.*, 14, 9855-9869, doi:10.5194/acp-14-9855-2014.

Cooper, O. R., D. D. Parrish, Ziemke, J., **N. V. Balashov**, M. Cupeiro, I. E. Galbally, S. Gilge, L. Horowitz, N. R. Jensen, J.-F. Lamarque, V. Naik, S. J. Oltmans, J. Schwab, D. T. Shindell, A. M. Thompson, V. Thouret, Y. Wang, and R. M. Zbinden (2014), Global distribution and trends of tropospheric ozone: an observation-based review, *Elem. Sci. Anth.*, 2, 10000029, doi:10.12952/journal.elementa.