Probabilistic Forecasting of Surface Ozone with a Novel Statistical Approach

NIKOLAY V. BALASHOV
Department of Meteorology, The Pennsylvania State University, University Park, Pennsylvania

ANNE M. THOMPSON
Department of Meteorology, The Pennsylvania State University, University Park, Pennsylvania, and Earth Sciences Division, NASA Goddard Space Flight Center, Greenbelt, Maryland

GEORGE S. YOUNG
Department of Meteorology, The Pennsylvania State University, University Park, Pennsylvania

(Manuscript received 10 March 2016, in final form 19 October 2016)

ABSTRACT

The recent change in the Environmental Protection Agency’s surface ozone regulation, lowering the surface ozone daily maximum 8-h average (MDA8) exceedance threshold from 75 to 70 ppbv, poses significant challenges to U.S. 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 U.S. AQ forecasters, this study explores a surface ozone MDA8 forecasting tool that is based solely on statistical methods and standard meteorological variables from the numerical weather prediction (NWP) models. The model combines the self-organizing map (SOM), which is 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 that are 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 on the basis of the weather patterns predicted by an NWP model. REGiS is evaluated over the San Joaquin Valley in California and the 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.

1. Introduction

The Environmental Protection Agency (EPA) designates surface ozone as a harmful pollutant and has recently reduced the ozone daily maximum 8-h average (MDA8) exceedance threshold from 75 to 70 ppbv across the United States, making the job for air quality (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 U.S. 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 United States also 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). To improve the U.S. ozone prediction capability, 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). These and related studies (e.g., Garner and Thompson 2013; Djalalova et al. 2015) have generally applied meteorological model ensembles or postprocessing techniques to chemical
transport models (CTMs) to quantify the uncertainty of AQ forecasts. A concise summary of these studies can be found in Zhang et al. (2012b). Although some of these approaches 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 CTMs that can also produce surface ozone probabilistic forecasts is based on statistical AQ modeling (Zhang et al. 2012a). Such an AQ statistical model would greatly benefit AQ forecasters as its predictions could be readily produced for most of the AQ monitoring stations in the United States, informing the forecasters on the range of likely ozone MDA8 outcomes. Combining CTMs and statistical AQ probabilistic ozone MDA8 forecasts would allow U.S. AQ managers to make robust decisions by taking into account the forecast uncertainty (Garner and Thompson 2012, 2013).

The most common statistical AQ approaches include classification and regression trees (CARTs), linear regression, artificial neural networks (ANNs), extreme value approaches, fuzzy logic, and Kalman filters (Burrows et al. 1995; Van der Wal and Janssen 2000; Perez and Reyes 2006; Cobourn 2007; Shad et al. 2009). These models rely on the fact that the air pollutants, such as surface ozone and particulate matter (PM), partially correlate with meteorology (Lelieveld and Crutzen 1990; Sillman and Samson 1995; Dueñas et al. 2002; Balashov et al. 2014). The first steps require an analysis of historical air pollutant and meteorological data to determine that relationship. Then, the relationship can be used to predict future AQ at a designated location based on the expected meteorological conditions. The advantage of such statistical modeling is that it offers moderate to high accuracy at a moderate cost (Zhang et al. 2012a). Certain statistical methods have been shown to perform with significant skill and in some cases even outperformed CTMs (Dutot et al. 2007). 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). To compensate for these weaknesses, a combination of different statistical tools working together could be used (Diaz-Robles et al. 2008). This study presents a statistical model that combines two statistical approaches to generate a probabilistic ozone MDA8 forecast that quantifies the uncertainty that is inherent in AQ prediction.

This hybrid statistical model combines the 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 here to evaluate local weather effects on ozone variability (Demuzere and van Lipzig 2010). 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 are organized with respect to their 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 section 4). 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 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 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 ANN. Besides serving as additional, uncertainty quantifying guidance for U.S. 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 article is to explain REGiS mechanics and to test the feasibility of the REGiS design. In this work, 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 operational forecasting this would not be the case and the meteorological data would have to be supplied by NWP systems such as the 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 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 subsequent sections, the foundation, implementation, tuning, and performance of REGiS are explored in detail. The model is first developed and is then evaluated using a couple of skill metrics for probabilistic forecasts exploring ozone data from the 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: a long-term station and a short-term site from the Deriving Information on Surface Conditions from Column and Vertically Resolved Observations Relevant to Air Quality (DISCOVER-AQ; Crawford and Pickering 2014) field campaign. The advantages and drawbacks of the approach are discussed. In section 2 the training and testing data are presented. Section 3 methodically describes the foundations and workings of the model. Section 4 explains the operational aspects of the model and how it is tuned for optimal performance. The results of this research are presented in section 5, where the model is evaluated using several sets of independent data. The last and concluding section summarizes this work.

2. Data

a. REGiS: Training and tuning versus testing

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, tuning of REGiS is undertaken at a single AQ site in the SJV (Fig. 1b), the Parlier site, using four different training periods during June–August (JJA): 1987–2012, 1995–2012, 2000–12, and 2005–12. REGiS is uniquely tuned for each of the mentioned training periods using a JJA 2013–14 validation dataset. The use of 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 the United States over this time period.

To maintain independence and thus reduce overfitting, the Parlier AQ monitoring station is not included in the final evaluation of REGiS performance. Table 1 summarizes the AQ and corresponding meteorological stations used in the REGiS evaluation process. Fresno Air Terminal (KFAT) data are used for the Clovis, Fresno–Drummond (FD), Fresno–SSP, and Parlier stations in California. Visalia Municipal Airport (KVIS) data are used for Hanford and Visalia–North Church, California. Meadows Field Airport (KBFL), also in California, is used for Oldale and Shafter. Greeley–Weld County Airport (KGXY) in Colorado is used for Greeley–Weld County Tower (WTC) and Platteville. Subsequent REGiS training for the independent AQ sites, shown in Table 1, 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. The results of these tests are evaluated in section 5.

b. Data for self-organizing maps

To distinguish between ozone-relevant synoptic-scale weather patterns using the SOM method, daily 500-hPa geopotential height $H_{500}$, 2-m $T$, 2-m $T_d$, 10-m $V$, 10-m $U$, and 850-hPa $T$ fields on a $0.75^\circ \times 0.75^\circ$ latitude–longitude grid at 0000 UTC are utilized from the ERA-Interim reanalysis (Dee et al. 2011). The data are generated by the European Centre for Medium-Range Weather Forecasts (ECMWF) and are available free of charge online (http://apps.ecmwf.int/datasets/). These variables are selected because they capture transport ($H_{500}$ and 10-m wind components) and stability ($H_{500}$, 850-hPa $T$, 2-m $T$, and 2-m $T_d$), 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 this study are centered on the states of California and Colorado (Fig. 1). 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 $H_{500}$ fields (Figs. 1a,c) are slightly larger than that for the other five variables (indicated by green rectangles in Fig. 1). Note that since REGiS is trained on historical analyses rather than historical NWP forecasts, it is classified as a perfect-prog method (Marzban et al. 2006).

c. Data for regression

Once the weather regimes are grouped by SOM, as described in section 3, a stepwise weighted quadratic regression equation is developed for each regime for an AQ site of interest. The idea is that each equation will correspond to a particular meteorological setting, better capturing ozone-meteorology sensitivity given a specific weather regime. Surface ozone MDA8 is a predictand in the regression. The predictor variables presently used in the regression model are listed in Table 2 and are considered standard in statistical AQ modeling, as suggested by Thompson et al. (2001).

Production of surface ozone is significantly affected by meteorology (Seinfeld and Pandis 2012). This fact allows for ozone MDA8 prediction using meteorological variables shown in Table 2. Surface temperature affects the rate of chemical reactions and is 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). A dewpoint temperature predictor is used as a proxy for atmospheric moisture content that alters hydroxyl radical concentrations, where wetter conditions tend to reduce ozone while drier 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 previous-day ozone.
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 the generation of ozone (Seinfeld and Pandis 2012). In summer, the zenith angle is smaller and more energy is available for NO2 photolysis, but in fall and spring the zenith angle is larger and less energy reaches the surface of the earth. The zenith angle may potentially be an important predictor when the regression training period occurs between summer and fall or between spring and summer. Although a day in July and a day in September may experience similar high temperatures, the photolysis rate would differ and correspondingly affect ozone production.

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 a 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 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 the current work.

The data used to fit the regression equations (Table 2) come from an AQ station and a nearby meteorological station (Table 1). Figure 1 shows the maps with all of the stations that are used to evaluate REGiS in this work. Additionally, Table 1 provides more details regarding each station used. 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/).

3. Ozone forecasting model

REGiS is inspired by the tree-based and stratified models that are founded 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 the regression trees, meteorological regimes or patterns are identified by SOM (Kohonen 2013), a technique related to ANNs. Once the regimes are identified, a stepwise weighted quadratic regression equation for predicting ozone MDA8 is developed for each weather pattern.

### Table 1. The AQ stations used to evaluate REGiS and the corresponding meteorological (METEO) stations along with related basic information.

<table>
<thead>
<tr>
<th>Station</th>
<th>WBAN/AQS ID</th>
<th>Type</th>
<th>Lat (°)</th>
<th>Lon (°)</th>
<th>Elev (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>KFAT</td>
<td>93193</td>
<td>METEO</td>
<td>36.78</td>
<td>−119.72</td>
<td>100</td>
</tr>
<tr>
<td>Clovis</td>
<td>06–019–5001</td>
<td>AQ</td>
<td>36.82</td>
<td>−119.72</td>
<td>86</td>
</tr>
<tr>
<td>Fresno–Drummond</td>
<td>06–019–0007</td>
<td>AQ</td>
<td>36.70</td>
<td>−119.74</td>
<td>89</td>
</tr>
<tr>
<td>Fresno-SSP</td>
<td>06–019–0242</td>
<td>AQ</td>
<td>36.84</td>
<td>−119.87</td>
<td>65</td>
</tr>
<tr>
<td>Parlier</td>
<td>06–019–4001</td>
<td>AQ</td>
<td>36.60</td>
<td>−119.50</td>
<td>78</td>
</tr>
<tr>
<td>KVIS</td>
<td>93144</td>
<td>METEO</td>
<td>35.32</td>
<td>−119.39</td>
<td>90</td>
</tr>
<tr>
<td>Hanford</td>
<td>06–031–1004</td>
<td>AQ</td>
<td>35.31</td>
<td>−119.64</td>
<td>99</td>
</tr>
<tr>
<td>Visalia–North Church</td>
<td>06–107–2002</td>
<td>AQ</td>
<td>35.33</td>
<td>−119.29</td>
<td>97</td>
</tr>
<tr>
<td>KBFL</td>
<td>23155</td>
<td>METEO</td>
<td>35.44</td>
<td>−119.05</td>
<td>151</td>
</tr>
<tr>
<td>Oildale*</td>
<td>06–029–0232</td>
<td>AQ</td>
<td>35.50</td>
<td>−119.27</td>
<td>126</td>
</tr>
<tr>
<td>Shafter**</td>
<td>06–029–6001</td>
<td>AQ</td>
<td>40.44</td>
<td>−104.63</td>
<td>1432</td>
</tr>
<tr>
<td>KGXY</td>
<td>24051</td>
<td>METEO</td>
<td>40.39</td>
<td>−104.74</td>
<td>1484</td>
</tr>
<tr>
<td>Greeley–WTC</td>
<td>08–123–0009</td>
<td>AQ</td>
<td>40.18</td>
<td>−104.73</td>
<td>1522</td>
</tr>
<tr>
<td>Platteville</td>
<td>—</td>
<td>AQ</td>
<td>40.18</td>
<td>−104.73</td>
<td>1522</td>
</tr>
</tbody>
</table>

* Full name of the Oildale AQ station is 3311 Manor St., Oildale, CA 93308.
** Full name of the Shafter AQ station is 548 Walker St., Shafter, CA 93263.

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

<table>
<thead>
<tr>
<th>Variable</th>
<th>Time scale</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zenith angle</td>
<td>Daily at 1200</td>
<td></td>
</tr>
<tr>
<td>Surface temperature</td>
<td>Mean over 1000–1800</td>
<td>°C</td>
</tr>
<tr>
<td>Dewpoint temperature</td>
<td>Mean over 1000–1800</td>
<td>°C</td>
</tr>
<tr>
<td>Wind direction</td>
<td>Mean over 1000–1800</td>
<td>°</td>
</tr>
<tr>
<td>Wind speed</td>
<td>Mean over 1000–1800</td>
<td>m s⁻¹</td>
</tr>
<tr>
<td>Sky cover</td>
<td>Mean over 1000–1800</td>
<td>okta</td>
</tr>
<tr>
<td>Previous-day ozone</td>
<td>Previous-day MDA8</td>
<td>ppbv</td>
</tr>
</tbody>
</table>
a. Self-organizing maps

In meteorology, SOM has been shown to be an effective tool for clustering similar cases (Hewitson and Crane 2002; Agarwal and Skupin 2008; Pearce et al. 2011; Jensen et al. 2012). An important feature of SOM, which distinguishes it from the other clustering algorithms such as k-means or hierarchical, is its ability to arrange a 2D lattice of clusters (called nodes) according to their similarities, rather than simply grouping data cases into randomly positioned categories (Agarwal and Skupin 2008). This metagrouping is accomplished by using a neighborhood function \( h_{i,k}(t) \) [shown in (4)], which allows nodes positioned next to each other in the lattice (neighbors) to influence each other’s centroid. This component of SOM provides a training advantage because it allows each node to be defined by a larger sample of training cases. REGiS utilizes this feature when generating its ozone PDF forecast.

The size of the node lattice (aka map) must be predetermined. Because results may be sensitive to the size and arrangements of nodes in the map, the dimensions of the node map (SOMdim) have been one of the most frequently discussed topics in the SOM literature (Kohonen 2013). This question does not have a “right” answer because the optimal dimensions of an SOM are determined by an application. Depending on the problem at hand, a compromise has to be found between the resolution and the statistical accuracy of the map (Kohonen 2013; Stauffer et al. 2016). In this work, maps of various sizes are tested to identify SOMdim, producing the most reliable forecasts on the tuning dataset.

To begin clustering, SOM nodes must be initialized before being iteratively trained. The nodes are linearly initialized along the dimensions of the SOM map through a uniform sampling from a subspace spanned by the first and second principal components of the input data (Johnson et al. 2008). After the initialization, REGiS uses a batch training algorithm to perform the unsupervised learning procedure to organize a map (Vesanto et al. 2000). This process is iterative, where at each step all of the training cases are partitioned into \( V \) groups (one for each node). The assignment of each training case to a particular node is based on minimizing the Euclidian distance between training case \( j \), represented by vector \( \mathbf{x}_j \), and the reference node \( i \), represented by vector \( \mathbf{m}_i \):

\[
\| \mathbf{x}_j - \mathbf{m}_i \| = \min_{i} \{ \| \mathbf{x}_j - \mathbf{m}_i \| \}, \quad (1)
\]

where index \( c \) denotes the best-matching unit, that is, the node that is closest to the training case \( \mathbf{x}_j \). Once the training cases are mapped to the nodes, the reference nodes are updated as a neighborhood (i.e., similarity) weighted average of the training cases:

\[
\mathbf{m}_i(t+1) = \frac{\sum_{k=1}^{m} h_{i,k}(t) s_k(t)}{\sum_{k=1}^{m} n_k h_{i,k}(t)}, \quad (2)
\]

where \( h_{i,k}(t) \) is the neighborhood function, \( t \) is an epoch or a count of iterations, \( n_k \) is a number of training cases in node \( k \), and \( s_k(t) \) is the sum of these training cases, where each training case of node \( k \) is denoted by \( \mathbf{x}_p \), as

\[
s_k(t) = \sum_{p=1}^{n_k} \mathbf{x}_p. \quad (3)
\]

The neighborhood function, \( h_{i,k}(t) \), 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 \( i \) in the lattice. Different neighborhood functions are available. Here, the Epanechnikov neighborhood function is chosen because it was shown to have the lowest quantization error when compared with other standard neighborhood functions (Liu et al. 2006; Stauffer et al. 2016). The Epanechnikov neighborhood function has the following mathematical representation:

\[
h_{i,k}(t) = \max\left\{ 0, 1 - \left( \frac{\| r_i - r_k \|}{\sigma(t)} \right)^2 \right\}, \quad (4)
\]

where \( \| r_i - r_k \| \) is the distance between units \( i \) and \( k \) on the SOM grid and \( \sigma(t) \) is the neighborhood radius at iteration \( t \). During the training, \( \sigma(t) \) decreases linearly with each repetition \( t \) until it becomes 1. The training begins 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 forms, 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. This feature is exploited by REGiS in the weighting of node-specific regression models to produce a forecast PDF. The SOM implementation used here is the SOM toolbox for MATLAB developed and written by the research group from the Laboratory of Information and Computer Science at the Helsinki University of Technology (available online at http://www.cis.hut.fi/somtoolbox/).

b. Regression in SOM

The construction of the REGiS forecasting system is based on the two core steps: first identifying
distinct synoptic patterns (via SOM) and then fitting the corresponding regression models (for every SOM node). The daily $H_{500}$, 2-m $T$, 2-m $T_d$, 10-m $V$, 10-m $U$, and 850-hPa $T$ fields from ERA-Interim are used to define the meteorological regime and serve as the input data for the SOM training algorithm. Because all of these variables have different magnitudes, they need to be standardized for fair comparison during the SOM procedure. The SOM process is illustrated in Fig. 2. In steps 1 and 2 in Fig. 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 adjusts based on the array similarity (for details on this training process, see section 3a). Last, step 5 shows the final SOM after the training cycle is finished.

Once the regimes are established (Fig. 3), a stepwise weighted quadratic regression equation is developed twice (with and without the previous-day ozone predictor; see section 2c) for each regime using the local meteorological and chemical variables as predictors (shown in Table 2) for ozone MDA8 (Comrie 1997). 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., overfitting). 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 the 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). For each regime, ozone MDA8 data are linearly detrended and adjusted to the latest observation point within that regime (see the appendix) in order to partially account for the changes in emissions over the years (Jhun et al. 2015).

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
The weights $v_{j,i}$ of the data at a node $j$, for the regression centered on the node $i$ are calculated as follows:

$$v_{j,i} = \frac{1}{\sum_{i=1}^{m} \left( \frac{1}{d_{j,i}} \right)^{\beta}} h_{i,j}, \quad (5)$$

where $\alpha$ and $\beta$ are the tuning parameters, $d_{j,i}$ is the Euclidean distance between the node centroids, and $h_{i,j}$ is the neighborhood function defined in (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 $\sigma$ determines the extent to which training cases from the surrounding nodes are used. When $\alpha \to \infty$ and $\beta \to 0$, 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 of the training cases.

**FIG. 3.** The $4 \times 6$ SOM of daily $H_{500}$ (m) from JJA over 1987–2012, centered on the western United States (domain shown in Fig. 1a). The daily fields of 2-m $T$, 2-m $T_d$, 10-m $V$, 10-m $U$, and 850-hPa $T$ are also used in this SOM training process.
4. Tuning and verification

a. Forecasting procedure

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 the NWP model) for a time of interest. The steps in this process are identification of the forecast weather pattern, assessment of its similarity to each of the SOM nodes, and the combination of the regression forecasts from these nodes to produce an ozone PDF forecast. The process is summarized in Fig. 4.

Recall that the weather pattern (i.e., SOM node) is determined based on the spatial distribution of several meteorological variables on the two domains. 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 an example of identified SOM weather patterns, see Fig. 3) using the 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 an SOM pattern. This is accomplished by using a generalization of the Cressman weighting function (Cressman 1959):

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

where $R_c$ is an adjustable radius of influence around the predicted pattern and $\gamma$ is a tuning parameter (Fig. 5). Note that $W$ is set to zero if $d > R_c$. In other words, $R_c$ determines the cutoff threshold beyond which the regression equation of an SOM pattern is not considered in making a prediction of the ozone PDF. Fine-tuning $R_c$ and $\gamma$ allows for various weight configurations, which influence the forecast probability distributions of ozone MDA8. These two parameters must, therefore, be tuned.
in order for REGiS to produce reasonably well-calibrated probabilistic forecasts of ozone MDA8.

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

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

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 the modeled kernel shapes at each of the values. For \(K\), a Gaussian smoothing kernel is used here because in this way the distribution tails are not cut off abruptly so more extreme scenarios are considered. Kernel density is sensitive to \(h\), which is estimated using the optimizing algorithm described in Bowman and Azzalini (1997).

b. Tuning

Tuning REGiS consists of finding optimal values for the parameters \(\alpha, \beta, \sigma, \gamma, R_c,\) and \(SOM_{\text{dim}}\) in the above-described model using a dataset independent of the one

---

**FIG. 5.** Diagram illustrating the Cressman radius \(R_c\), which is set by a user and is the influence radius of the predicted weather pattern \(i\) on the weights of the SOM weather patterns \(m\). It takes a value of some Euclidian distance \(d_{mi}\) (distance between patterns \(i\) and \(m\)) from \(d_{mi} = 0\) to \(\max(d_{mi})\).

**FIG. 6.** Example of the REGiS tuning process at the Parlier AQ station. In the given example, the training data period is JJA 1987–2012 and the validation data period is JJA 2013–14.
used to train the SOM and the regression models. The goal here is to ensure that the PDF forecast represents the true distribution of a predicted variable. In this work, the verification rank histogram (Hamill 2001) is used for such a purpose: to quantify the reliability of a forecasting system. 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 observations may fall. The predicted PDF is statistically indistinguishable from the verifying observations 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

\[
\text{RMSD} = \sqrt{\frac{1}{N+1} \sum_{k=1}^{N_1} \left( s_k - \frac{M}{N+1} \right)^2},
\]

where RMSD is the root-mean-square deviation from complete 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 the value of RMSD is, the flatter is the corresponding rank histogram. The idea is to tune REGiS parameters in such 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 \(a\), \(b\), \(s\), \(R_c\), and \(g\) 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 Fig. 6. The RMSD is determined by using the validation data over JJA 2013–14. To begin, an educated guess is made for the six parameters. Then, for \(s = 2\) and \(s = 3\), \(a\) and \(b\) are tested. For \(s = 1\), \(a\) and \(b\) make no impact. The lowest RMSD occurs when \(s = 2\), \(a = 0\), and \(b = 0.5\). Next, \(g\) is adjusted, while all of the other parameters are held constant. Minimum RMSD happens when \(g = 0\). Using the same approach \(R_c\) is determined to
be 100. Finally, iteration over different SOM$_\text{dim}$ is performed.

Four tuning processes are completed in this work using the procedure described above for the four different time periods using the Parlier data. Figure 7 shows the results of the final tuning step. The color plot shows RMSDs for probabilistic daily REGiS forecasts (for JJA 2013–14) given various SOM$_\text{dim}$. The color bar represents RMSD in terms of the number of cases (observation points). Figure 8 presents 10 of the best REGiS SOM configurations from Fig. 7 sorted by RMSD. The results indicate that, in theory, using the training periods JJA 1995–2012 and JJA 2000–12 should yield better REGiS performance than when using the other two periods.

c. Verification

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

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

where

$$F_0(y) = \begin{cases} 0, & y < \text{observed value} \\ 1, & y \geq \text{observed value} \end{cases} \quad (9b)$$

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 the predicted PDF. In this work, CRPS is used as an additional skill metric when comparing REGiS’s performance among different stations.

5. Results: Model performance

When REGiS completes its forecasting process, the product in the form of a PDF is generated. Figure 9 demonstrates a sample product produced by REGiS: the
PDF of the ozone MDA8 forecast for the FD AQ station on 11 June 2014. In Fig. 9, ozone MDA8 runs along the x axis and the probability density is shown along the y axis. The black line indicates PDF, the area of which adds up to one. The shaded areas inside the PDF are the color-coded ozone pollution air quality index (AQI) categories, converted to ozone mixing ratios (ppbv) from AQI as designated by EPA (EPA 2015, 2016), where green, yellow, orange, and red indicate the 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 top-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 the moderate category.

To evaluate REGiS, the procedure described above has been carried out for the summers (JJA) of 2013 and 2014 at AQ sites in SJV and over the northeastern plains of Colorado (Table 3). This evaluation period is independent from that used for the training (see section 2a for more details). Different training periods are tested at each station, while the tuning parameters determined in section 4 for Parlier are used. Various training periods are examined because of constantly changing ozone precursor emissions at each site. Although in this study 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 linear 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’s 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 fifth and sixth columns in Table 3.

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 Fig. 10. The prediction is generated from the 1995–2012 training dataset (see Table 3). Figure 10a shows REGiS probabilistic forecasts in terms of percentiles, where gray and yellow colors highlight the 5th–95th and 25th–75th percentiles of the PDFs, respectively. The blue line denotes observations and the red dashed line is EPA’s exceedance threshold for ozone MDA8. Figure 10b displays the corresponding rank histogram, revealing into which percentile regions the observations fall across the PDFs. There are five 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 during JJA of 2013 and 2014, fall into the abovementioned five percentile regions. For station FD, the rank histogram is slightly skewed to the left (model is biased high), with an RMSD of 6.9 cases. The CRPS, which indicates the 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.

Further examination of Table 3 reveals that RMSD and CRPS generally increase (skill lowers) for the stations that are farther away from KFAT and Parlier, sites used to initially tune the REGiS parameters: $\alpha$, $\beta$, $\sigma$, $\gamma$, $R_c$, and $\text{SOM}_{\text{dim}}$. Specifically, the two 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 1). The RMSD results begin to increase as we apply REGiS to other AQ stations, where KVIS and KBFL are used as meteorological references. One potential reason for this increase is examined below by looking more closely at the Oildale AQ site.

Figure 11 presents the REGiS evaluation at Oildale for the two training periods. Oildale (b) REGiS (see Table 3) is developed using 1994–2012 JJA data in
Table 3. Summary of the REGiS evaluation at nine different stations over SJV and northeastern CO.

<table>
<thead>
<tr>
<th>Station</th>
<th>SOM dimensions</th>
<th>Training period</th>
<th>Evaluation period</th>
<th>CRPS (ppbv)</th>
<th>RMSD (No. cases)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clovis (a)</td>
<td>4 x 6</td>
<td>JJA 1995–2012</td>
<td>JJA 2013–14</td>
<td>4.95</td>
<td>12.0</td>
</tr>
<tr>
<td>Clovis (b)</td>
<td>7 x 7</td>
<td>JJA 2000–12</td>
<td>JJA 2013–14</td>
<td>4.71</td>
<td>10.4</td>
</tr>
<tr>
<td>Clovis (c)</td>
<td>5 x 3</td>
<td>JJA 2005–12</td>
<td>JJA 2013–14</td>
<td>4.71</td>
<td>12.0</td>
</tr>
<tr>
<td>FD (a)</td>
<td>6 x 11</td>
<td>JJA 1987–2012</td>
<td>JJA 2013–14</td>
<td>4.69</td>
<td>12.9</td>
</tr>
<tr>
<td>FD (b)</td>
<td>4 x 6</td>
<td>JJA 1995–2012</td>
<td>JJA 2013–14</td>
<td>4.38</td>
<td>6.9</td>
</tr>
<tr>
<td>FD (c)</td>
<td>7 x 7</td>
<td>JJA 2000–12</td>
<td>JJA 2013–14</td>
<td>4.63</td>
<td>11.5</td>
</tr>
<tr>
<td>FD (d)</td>
<td>5 x 3</td>
<td>JJA 2005–12</td>
<td>JJA 2013–14</td>
<td>5.46</td>
<td>21.8</td>
</tr>
<tr>
<td>Fresno-SSP (a)</td>
<td>6 x 11</td>
<td>JJA 1987–2012</td>
<td>JJA 2013–14</td>
<td>4.70</td>
<td>19.8</td>
</tr>
<tr>
<td>Fresno-SSP (b)</td>
<td>4 x 6</td>
<td>JJA 1995–2012</td>
<td>JJA 2013–14</td>
<td>4.03</td>
<td>5.6</td>
</tr>
<tr>
<td>Fresno-SSP (c)</td>
<td>7 x 7</td>
<td>JJA 2000–12</td>
<td>JJA 2013–14</td>
<td>4.45</td>
<td>15.8</td>
</tr>
<tr>
<td>Fresno-SSP (d)</td>
<td>5 x 3</td>
<td>JJA 2005–12</td>
<td>JJA 2013–14</td>
<td>4.45</td>
<td>14.7</td>
</tr>
<tr>
<td>Hanford (a)</td>
<td>4 x 6</td>
<td>JJA 1994–2012</td>
<td>JJA 2013–14</td>
<td>4.32</td>
<td>7.7</td>
</tr>
<tr>
<td>Hanford (b)</td>
<td>7 x 7</td>
<td>JJA 1998–2012</td>
<td>JJA 2013–14</td>
<td>4.68</td>
<td>10.6</td>
</tr>
<tr>
<td>Hanford (c)</td>
<td>5 x 3</td>
<td>JJA 2003–12</td>
<td>JJA 2013–14</td>
<td>4.21</td>
<td>7.7</td>
</tr>
<tr>
<td>Visalia–North Church (a)</td>
<td>6 x 11</td>
<td>JJA 1985–2012</td>
<td>JJA 2013–14</td>
<td>5.34</td>
<td>23.6</td>
</tr>
<tr>
<td>Visalia–North Church (b)</td>
<td>4 x 6</td>
<td>JJA 1995–2012</td>
<td>JJA 2013–14</td>
<td>5.20</td>
<td>28.9</td>
</tr>
<tr>
<td>Visalia–North Church (c)</td>
<td>7 x 7</td>
<td>JJA 2000–12</td>
<td>JJA 2013–14</td>
<td>6.10</td>
<td>32.3</td>
</tr>
<tr>
<td>Visalia–North Church (d)</td>
<td>5 x 3</td>
<td>JJA 2005–12</td>
<td>JJA 2013–14</td>
<td>5.14</td>
<td>22.5</td>
</tr>
<tr>
<td>Oildale (a)</td>
<td>6 x 11</td>
<td>JJA 1986–2012</td>
<td>JJA 2013–14</td>
<td>6.30</td>
<td>33.2</td>
</tr>
<tr>
<td>Oildale (b)</td>
<td>4 x 6</td>
<td>JJA 1994–2012</td>
<td>JJA 2013–14</td>
<td>6.33</td>
<td>42.1</td>
</tr>
<tr>
<td>Oildale (c)</td>
<td>7 x 7</td>
<td>JJA 1999–2012</td>
<td>JJA 2013–14</td>
<td>6.10</td>
<td>35.6</td>
</tr>
<tr>
<td>Oildale (d)</td>
<td>5 x 3</td>
<td>JJA 2004–12</td>
<td>JJA 2013–14</td>
<td>5.09</td>
<td>29.2</td>
</tr>
<tr>
<td>Shafter (a)</td>
<td>4 x 6</td>
<td>JJA 1995–2012</td>
<td>JJA 2013–14</td>
<td>5.08</td>
<td>22.8</td>
</tr>
<tr>
<td>Shafter (b)</td>
<td>7 x 7</td>
<td>JJA 2000–12</td>
<td>JJA 2013–14</td>
<td>5.09</td>
<td>17.8</td>
</tr>
<tr>
<td>Shafter (c)</td>
<td>5 x 3</td>
<td>JJA 2005–12</td>
<td>JJA 2013–14</td>
<td>4.62</td>
<td>13.0</td>
</tr>
<tr>
<td>Greeley–WTC (a)</td>
<td>3 x 7</td>
<td>JJA 2002–12</td>
<td>JJA 2013–14</td>
<td>3.61</td>
<td>7.9</td>
</tr>
<tr>
<td>Greeley–WTC (b)</td>
<td>5 x 3</td>
<td>JJA 2005–12</td>
<td>JJA 2013–14</td>
<td>3.71</td>
<td>8.5</td>
</tr>
<tr>
<td>Platteville (a)</td>
<td>3 x 7</td>
<td>JJA 2002–12</td>
<td>JA 2014</td>
<td>3.85</td>
<td>4.7</td>
</tr>
<tr>
<td>Platteville (b)</td>
<td>5 x 3</td>
<td>JJA 2005–12</td>
<td>JA 2014</td>
<td>3.78</td>
<td>4.7</td>
</tr>
</tbody>
</table>

a Years 2008 and 2009 are missing.
b Year 2005 is missing.
c Here, JA stands for 13 Jul–11 Aug.

Using the settings determined at Parlier, REGiS is also evaluated for northeastern Colorado at Greeley–WTC, a long-term AQ station, and a short-term DSICOVER-AQ campaign site at Platteville (Crawford and Pickering 2014). Platteville, which is located about 30 km south of Greeley–WTC (Fig. 1), is used to test whether the REGiS ozone MDA8 predictions for an AQ station (in this case Greeley–WTC) are representative of its surroundings with available ozone data from 13 July to 11 August 2014. For both stations, REGiS is trained using the KGXY meteorological station and the Greeley–WTC AQ site. At Greeley–WTC, RMSD and CRPS values are comparable to those at the sites in the SJV (Table 3), indicating that REGiS is able to perform in Colorado at least as well as it does in the 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 examination at both of the stations suggests that the REGiS

contrast to Oildale (d) REGiS, which is developed using 2004–12 JJA data. From Figs. 11a,b it is possible to notice that the 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 value in this study (Fig. 11b). 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 11c,d still indicate 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 JJA 2004–12 seem to be more representative of the ozone MDA8 observations during JJA of 2013 and 2014. So, although the REGiS tuning process at Parlier implied that the best training periods are 1995–2012 and 2000–12, that may not always be the case, as the current example with Oildale AQ site illustrates.
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 dataset. And 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 section 4. Various sets of training data could be tested, but it is recommended to use at least 4 yr of data to make sure that the sample size is sufficient for SOM’s clustering procedure. It is best to train REGiS on a specific season, such as here where the summer data are used. Similarly, the data could be separated into the other seasons: autumn, winter, and spring. The meteorological stations that supply predictors of MDA8 ozone for REGiS should be supported by MOS products to allow REGiS to run operationally.

In the presented REGiS testing experiment, it is assumed that the meteorology is known perfectly and hence ERA-Interim and NCDC observations are used to make the 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. In this study only 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 the reliable distribution of the potential MDA8 ozone mixing ratios. By itself, REGiS may not provide an AQ forecaster with precise 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.

6. Conclusions

In this study, a novel statistical approach has been developed to produce probabilistic daily surface ozone MDA8 forecasts. The model draws its inspiration from 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. The 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, as well as 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 following three words: regression in SOM.

REGiS is evaluated at SJV, a location known for its poor AQ, and in northeastern Colorado 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 KFAT meteorological station. Four training periods are used to study REGiS’s sensitivity to the change in emissions. Once tuned, REGiS is tested at 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 the flat rank histogram. CRPS determines the 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 the regressions, REGiS achieves lower RMSD values than for the stations that are farther away and that used other meteorological sites for the regressions. CRPS behaves similarly, with the exception of the Hanford AQ station. This implies that tuning REGiS for each station separately would generate a more robust model for probabilistic MDA8 ozone prediction.

The uniqueness of REGiS is in its ability to generate probabilistic MDA8 ozone forecasts. Uncertainty quantification using a PDF gives an advantage to probabilistic forecasts over discrete deterministic forecasts. 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 setup. This work underlines the value of the REGiS ozone MDA8 probabilistic forecasts and their ability to aid AQ forecasters by quantifying the uncertainty of operational ozone prediction.

Acknowledgments. The authors thank three reviewers for their numerous valuable comments that helped considerably to improve this article. The authors acknowledge the Seventh International Workshop on Air Quality Forecasting Research (IWAQFR), held in College Park, Maryland, in 2015, for important input regarding the state of international AQ forecasting in 2015. Special thanks are given to AQ operational forecasters William Ryan, Amy Huff, and Joel Dreessen for their feedback regarding the AQ forecasting work flow. Thanks are also given to NASA for their enlightening DISCOVER-AQ campaign experience. We give many thanks to Ryan Stauffer, Hannah Haliday, Greg Garner, and the rest of the Gator research group for their continuing help and support. This research was supported by NASA through grants to The Pennsylvania State University, DISCOVER-AQ (NNX10AR39G), and Applied Sciences Air Quality (NNX11AQ44G).

APPENDIX

Detrending Ozone MDA8 Values

To partially account for the changes in ozone precursors emissions over the years (Jhun et al. 2015), ozone MDA8 data are detrended after the data have...
been separated into unique regimes by the SOM (Fig. A1). 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 dataset (Fig. A2a). Once the detrending is accomplished, the entire detrended MDA8 time series is adjusted up by the difference between the last value in the original MDA8 ozone dataset and the last value in the detrended MDA8 ozone dataset to generate the final detrended dataset (Fig. A2b).

REFERENCES


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


