Ensemble and bias-correction techniques for air-quality model forecasts of surface O\textsubscript{3} and PM\textsubscript{2.5} during the TEXAQS-II experiment of 2006.

I.Djalalova\textsuperscript{1,2}, J.Wilczak\textsuperscript{2}, S.McKeen\textsuperscript{1,3}, G.Grell\textsuperscript{1,4}, S.Peckham\textsuperscript{1,4}, M.Pagowski\textsuperscript{1,4}, L.DelleMonache\textsuperscript{5}, J.McQueen\textsuperscript{6}, P.Lee\textsuperscript{6}, Y.Tang\textsuperscript{6}, J.McHenry\textsuperscript{7}, W.Gong\textsuperscript{8}, V.Bouchet\textsuperscript{9}, R.Mathur\textsuperscript{10}

\textsuperscript{1}Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado, Boulder, Colorado, USA

\textsuperscript{2}Earth System Research Laboratory/Physical Sciences Division, National Oceanic and Atmospheric Administration, Boulder, Colorado, USA

\textsuperscript{3}Earth System Research Laboratory/Chemical Sciences Division, National Oceanic and Atmospheric Administration, Boulder, Colorado, USA

\textsuperscript{4}Earth System Research Laboratory/Global Systems Division, National Oceanic and Atmospheric Administration, Boulder, Colorado, USA

\textsuperscript{5}National Center for Atmospheric Research, Boulder, Colorado, USA

\textsuperscript{6}National Weather Service/National Center for Environmental Prediction/Environmental Modeling Center, National Oceanic and Atmospheric Administration, Camp Springs, Maryland, USA

\textsuperscript{7}Baron Advanced Meteorological Systems, Raleigh, North Carolina, USA

\textsuperscript{8}Environment Canada, Science and Technology Branch, Downsview, Ontario, Canada

\textsuperscript{9}Environment Canada, Meteorological Service of Canada, Dorval, Quebec, Canada

\textsuperscript{10}Environmental Protection Agency/National Exposure Research Laboratory, Research Triangle Park, North Carolina, USA

Corresponding Author: I.Djalalova, 325 Broadway, Boulder, CO, 80303

Phone: 303-497-6238, FAX: 303-497-6181, e-mail: Irina.V.Djalalova@noaa.gov
Abstract:

Several air quality forecasting ensembles were created from seven models, running in real-time during the 2006 Texas Air Quality (TEXAQS-II) experiment. These multi-model ensembles incorporated a diverse set of meteorological models, chemical mechanisms, and emission inventories. Evaluation of individual model and ensemble forecasts of surface ozone and particulate matter (PM) was performed using data from 119 EPA AIRNow ozone sites and 38 PM sites during a 50-day period in August and September of 2006. From the original set of models, two new bias-corrected model data sets were built, either by applying a simple running-mean average to the past 7 days of data or by a Kalman-Filter approach. From the original and two bias-corrected data sets, three ensembles were created by a simple averaging of the seven models. For further improvements three additional weighted model ensembles were created, where individual model weights were calculated using the singular value decomposition method. All six of the ensembles are compared to the individual models and to each other in terms of root-mean-square error, correlation, and contingency and probabilistic statistics. In most cases, each of the ensembles show improved skill compared to the best of the individual models. The overall best ensemble technique was found to be the combination of Kalman-Filtering and weighted averaging. PM$_{2.5}$ aerosol ensembles demonstrated significant improvement gains, mostly because the original model’s skill was very low.

Keywords:

Air quality, ozone, particulate matter, TEXAQS 2006, ensemble forecast
1. Introduction.

Although widely used in operational weather prediction, ensemble forecasting of air quality has only begun to be investigated in recent years. Several different ensemble techniques have demonstrated better performance for ensemble forecasts compared to individual model forecasts for ozone, as discussed in Wilczak et al. (2006), McKeen et al. (2005, 2007), Vautard (2007, 2008), van Loon (2007), Pagowski (2005, 2006), Delle Monache et al. (2006a, 2006b), and others. We use surface ozone \( (O_3) \) and particulate matter (diameter less than 2.5 micrometers, \( PM_{2.5} \)) data from the Texas Air Quality Study (TexAQS-II) regional experiment of 2006 (Parrish, 2009) to show the benefits of the ensemble approach.

Observed \( O_3 \) and \( PM_{2.5} \) were compared with the seven models (with and without bias correction) in real-time during the experiment. These comparisons are available at [http://www.etl.noaa.gov/programs/2006/texas/verification/](http://www.etl.noaa.gov/programs/2006/texas/verification/), where observed and modeled meteorological and chemical variables, including \( O_3 \) and \( PM_{2.5} \), are shown on an hourly basis between August 01 and mid-October, 2006. At the time of the experiment, the modeling data appeared for 48 hours following the model initialization times, and the observation points were added hourly from 11 meteorological sites and 14 chemical sites within the Texas study region. In addition, real-time bias corrected ensemble \( O_3 \) and \( PM_{2.5} \) forecasts were shown.

In the present analysis we use data from a much larger set of sites that are part of the EPA AIRNOW observation network. This network includes 119 \( O_3 \) stations and 38 \( PM_{2.5} \) stations within Texas, as shown in Figure 1. The analysis period spans 50 days from August 12 through September 30, 2006. Seven models are included in the analysis: WRF/Chem version 2.2 using two horizontal resolutions of 12 and 36 km; the Canadian operational CHRONOS model (21 km horizontal grid) and the research AURAMS model (28 km horizontal grid), both provided by Environment Canada; the
NOAA/NWS operational CMAQ/NAM model run on a 12km horizontal grid; the 15 km resolution BAMS model (Baron Advanced Meteorological System, Inc.); and the STEM 12km horizontal grid forecast model (University of Iowa). A very detailed description of these models and their domain outlines may be found in McKeen et al. (2007, 2008), and a brief description is given in section 2.

Air quality models have been providing ozone forecasts for a number of years, and those forecast models now have relatively good skill at simulating the observed ozone variability. In contrast, PM$_{2.5}$ predictions by air quality models are relatively new and most forecast models have low forecast skill compared to ozone. From this perspective, it is of interest to compare the value of bias-correction and ensemble techniques for these two variables that have very different levels of predictive skill.

Section 3 contains a description of different ensemble techniques using O$_3$ data. We show not only time dependent comparisons, but also spatial comparisons of the ensemble and observations, as well as contingency and probability statistics. In section 4 we apply these techniques to PM$_{2.5}$, and a summary and conclusions are given in section 5.

2. Description of the models.

The seven models included in this analysis consist of both operational and research models. All of the model runs were provided in real-time during the experiment except the WRF-Chem simulations which were re-run to fill in extensive gaps in the real-time data set. Of the remaining models, some small data gaps exist due to incomplete model runs, and these were filled using the previous days forecast.

2.1 NOAA/ESRL/Global Science Division WRF-CHEM, version 2.2.

The Weather Research and Forecasting (WRF) Chemical model is based on the widely used WRF-ARW model (http://www.wrf-model.org/index.php) with “online” chemistry (Grell et al., 2005 and Fast et al., 2006). This is the only model in the analysis in which the meteorological and chemical fields are
calculated on the same grids and time steps. Meteorological initial conditions are taken from the RUC model, the chemistry is based upon the RADM2 (Stockwell et al., 1990, 1995) chemical mechanism, and emissions are taken from EPA 1999 National Emission Inventory (Frost et al., 2006). An inner grid of 12 km horizontal resolution is used, which is embedded in a 36 km domain with one-way nesting. Both the 36 km and 12 km models were initialized at 00Z and 12Z and ran for 36 hours.

2.2 Environment Canada CHRONOS and AURAMS models.

The Environment Canada CHRONOS model (http://www.weatheroffice.gc.ca/chronos/index_e.html; Pudykiewicz et al., 1997) has been used operationally since 2001. AURAMS (Gong et al., 2006; Smyth et al., 2008) is similar to CHRONOS, in that it was built upon the CHRONOS CTM, but has additional aerosol microphysics. Also, while both models run "offline" and use the Canadian Global Environmental Multiscale (GEM, Cote et al., 1998a;) forecast model to calculate the meteorological fields and the SMOKE anthropogenic emission system, CHRONOS is driven by meteorological fields updated at hourly intervals, whereas AURAMS uses meteorological fields updated every 15 minutes. Both models have the same forecast period of 48 hours, starting at 00Z.

2.3 Baron Advanced Meteorological Systems, Inc. (BAMS) MAQSIP-RT model.

MAQSIP-RT (McHenry et al., 2004) is an “off-line” chemical-transport model, applied for real-time ozone forecasting since 1998. Meteorological information is calculated through the WRF-ARW (version 2) model with initial and boundary conditions derived from NCEP’s operational GFS model. Chemical information is configured with a modified Carbon-Bond 4 (CBM-4b) chemical mechanism (Gery et al., 1989). During the field experiment the model was running on three horizontal grids, but we use only the 15 km resolution grid. All emissions components are computed within the SMOKE
(version 2.5-RT) system using the EPA NEI Version 3 (2001) anthropogenic emission inventory. The
model ran 48 hour forecasts starting at 06 UTC (used for this analysis) and 18Z UTC.

2.3 University of Iowa STEM-2K3 model.

STEM is a forecast model initially used for simulating sulfur dioxide (SO₂) and generalized later to
simulate regional air quality (http://rain.cger.uiowa.edu/TexAQS/texaqs-2l6.html). This model runs
“offline” using WRF-ARW meteorology with GFS boundary and initial conditions, the SAPRC-99
gaseous mechanism, and boundary conditions from the RAQMS global chemical transport model.
Emission components are calculated through the SMOKE system. STEM-2K3 has a 12 km horizontal
grid running a 48 hour forecast starting at 00Z UTC.

2.4 NOAA/NWS CMAQ/NAM (North American Mesoscale) model.

NMM-CMAQ is NOAA’s operational air quality model. During 2006 it was run in three different
modes: 1) operational O₃ forecasts over the eastern USA, 2) experimental O₃ forecasts over the
contiguous USA, and 3) developmental forecasts for both O₃ and PM₂.₅ over the contiguous USA. We
used the third real-time set of data. The model used a 12 km horizontal grid, with meteorological
fields provided by the NWS/NCEP WRF-NMM forecast, and photochemistry from the CMAQ model
(Byun and Schere, 2006). A conversion preprocessor, PREMAQ, is used to transform the WRF-NMM
meteorological fields into the CMAQ grid coordinates (Otte et al., 2005). The photochemical
mechanism within CMAQ is a modified CBM-4. The emission processing is based on the SMOKE
system with updated emission inventories to represent the 2006 forecast period. CMAQ/NAM
provided a 48 hour forecast starting at 06Z UTC.

Because of the different initialization times and forecast lengths of the various models, the time
window for the analysis was constrained to the 24-h period between 10-34 UTC for each days
forecast. All of the models in the ensemble were initialized at 00 UTC except for BAMS and NMM-CMAQ, which were initialized at 06 UTC.

It is seen from the brief model description that the different models use a variety of meteorological, chemical mechanism, and emission components that make possible a diverse multi-model ensemble. No post-processing was done by any of the model groups that provided data, so that we worked with raw model output. In some cases the model represents a single configuration of model parameters from a variety of possible choices. Finally, we note that since the field experiment, improvements to many of the models have been made, increasing their skill beyond that which is presented here.

3. Ensemble techniques and $O_3$ comparison.

As the first step of the evaluation, the observed $O_3$ and $PM_{2.5}$ values were matched with the closest model grid values. We chose to use a nearest point comparison rather than interpolation, to avoid smoothing during the interpolation; this allowed for the full dynamic range of model values to be represented. However, we note that without interpolation there were a few occurrences in which $O_3$ observation values from several monitors are compared against the same grid model value.

3.1 Bias-corrected ensembles.

Using hourly $O_3$ data values from the set of 7 input models, three different ensembles have been created. All three ensembles are formed as the simple averages of the models at each particular site and hour. The first ensemble (raw ensemble) uses the original models. Because all of the models are found to have significant biases (Figure 2), the other two ensembles average model values after bias corrections have been applied. We use two bias correction schemes, both of which use model data from the 7 previous days for each particular hour of the day. This 7 days training period was
chosen as a compromise between having a sufficiently long period to gather adequate statistics, but not too long to mask seasonal variations in ozone, as discussed in (Wilczak et al., 2006). In the first technique (7DRM_ensemble), we use just the simple running mean average of these 7 data points to correct the forecast day's model value. In the second case (KF_ensemble), we use the Kalman-Filter (KF) method (Brookner, 1998), again using the 7 previous days to estimate the bias between the original model and the observation. The KF is a recursive and adaptive method, because it predicts the future bias using a linear relationship between the previous bias estimate and the difference between the previous forecast error and the previous bias estimate. The KF is optimal in a least-square-error sense. A detailed description of KF as implemented in this analysis, is given in (Delle Monache et al., 2006a, 2006b). Because 7DRM_ensemble and KF_ensemble are both bias corrected ensembles, they match the observed ozone diurnal cycle almost perfectly (Figure 3). In comparison, the raw ensemble has a bias which reaches more than 10 ppbv in the morning hours, and 5 ppbv near the afternoon peak.

Following the Environment Protection Agency standard, our analysis is based on the daily 8-hour maximum ozone. The 8-hour maximum ozone is calculated for each day by first using a sliding window to produce a time-series of 8-hour averaged ozone for the observations, models, and ensembles, and then selecting the maximum 8-hour value in the 24 hour window corresponding to 10-34 hours from the start of the 00 UTC forecasts, both for observation and all models, which minimizes any ambiguity of the O$_3$ peak, as described in detail in Wilczak et al. (2008). The start and end times of this window correspond to the approximate minima in the mean observed ozone diurnal cycle (Figure 2). All of the observation data were visually inspected for potential errors, and all data were accepted except for one transient observed O$_3$ spike of 178 ppbv.

The 8-hour maximum O$_3$ time series for the observations and raw models for every site are shown in Figure 4(a). Also shown as a thin red line (for the observations) or black lines (for the
models) are the daily mean values of all sites. From this figure we can see that the observed values have a greater spatial variability in the early part of the field campaign, with noticeably less variability in the last two weeks, from days 35-50. Also apparent in the daily mean values is a pronounced approximately weekly cycle associated with synoptic weather events (weekday/weekend emissions variations do not fully explain the cycle, as ozone minima occur on almost all days of the week). To varying degrees the individual models replicate these features. The NMM-CMAQ and two WRF-Chem models are seen to have too little spatial variability and too small of a weekly variation. Both of these deficiencies can be explained by the erroneous ozone boundary conditions they use over the Gulf of Mexico, and by too small ozone precursor emissions in Houston, as discussed in Wilczak et al (2008). After applying the 7DRM and KF bias corrections (Figures 4b,c), the models are seen to better represent the spatial variability of ozone and its variation in time, especially for the KF corrections. The KF corrected data is also seen to most closely follow the observed weekly variation in 8-hour max ozone.

Using the raw and bias-corrected 8-hour maximum ozone values, we calculated bulk statistics parameters, such as bias, root mean square error and correlation coefficient (Figure 5). The solid bars represent the original models, the hatched bars represent the 7DRM models, and open bars represent the KF models. At the far right end of the plot, there are three types of ensembles: raw ensemble, 7DRM bias corrected ensemble and KF ensemble and their SVD modifications (explained later). The solid line is the climatology value and dashed line indicates the persistence forecast value.

The two bias correction schemes improve the statistics for all models, but larger improvements for most models occur with the KF scheme. Also, the raw ensemble has smaller rmse than all of the individual models, and higher correlation coefficient than all of the models except CHRONOS.

Further improvement is found with the 7DRM ensemble, but the best statistics come from the KF ensemble, which is also better than the all of the KF bias-corrected individual models as well, for
both rmse and correlation. The improvement of the KF ensemble over the raw ensemble is 26 percent for rmse and 9 percent for the correlation coefficient (Table 1).

3.2 SVD Ensembles.

The three ensembles considered so far equally weight all models, either original or bias corrected. A different approach to derive a bias-corrected ensemble is to use linear regression techniques to find weight coefficients for the models such that sum of the weighted models has a minimum bias (Krishnamurti et al., 1999, Pagowski, 2005).

To this end, using the hourly model and observed ozone values, we write the weighted model summation:

\[ M_{11}W_1 + M_{12}W_2 + ... + M_{17}W_7 + W_8 = O_i, \]  

(1)

where \( M_{11}, \ldots, M_{17} \) and \( O_i \) are the values of the 7 model forecasts and the observation from one of the previous 7 days for the same particular hour of the day for a single observation site; \( W_1, \ldots, W_7 \) are the unknown weight coefficients and \( W_8 \) is a bias. Generalizing (1) for all 7 previous days (similar to the 7 day training period for the 7DRM and KF bias corrections) and all AIRNOW network sites, we get an over-determined matrix system of \( I \sim 7 \times 100 = 700 \) equations like (1):

\[ M \times W = O. \]  

(2)

where

\[
M = \begin{bmatrix}
M_{11} & M_{12} & \cdots & M_{17} & 1 \\
M_{21} & M_{22} & \cdots & M_{27} & 1 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
M_{71} & M_{72} & \cdots & M_{77} & 1
\end{bmatrix}, \quad W = \begin{bmatrix}
W_1 \\
W_2 \\
\vdots \\
W_7 \\
W_8
\end{bmatrix}, \quad O = \begin{bmatrix}
O_1 \\
O_2 \\
\vdots \\
O_7
\end{bmatrix}
\]

To solve (2), we find \( M^{-1} \) using a least-squares minimization procedure called Singular Value Decomposition (SVD) which provides the optimal coefficients \( W_1, \ldots, W_7 \) and the bias \( W_8 \). The ensemble forecast for the present day is then calculated as:
SVD_Ensemble = M_{i1}^f * W_1 + M_{i2}^f * W_2 + ... + M_{i7}^f * W_7 + W_8, \quad (3)

where M_{i1}^f, M_{i2}^f, ..., M_{i7}^f are the hourly O_3 values of 7 models at the present forecast day. The same set of W_i coefficients and bias are used for all sites. This procedure is repeated for every hour of the day and for all days in the considered time period, and is applied to the raw, 7DRM and KF bias-corrected models, generating the new set of ensembles: SVD_ensemble, SVD_7DRM_ensemble and SVD_KF_ensemble.

The skill levels of the SVD ensembles and the previously discussed ensembles are shown in Fig. 5 and Table 1. Although the SVD ensemble is less skillful than the KF ensemble, the highest skill comes from the combination of both the SVD and KF techniques. Comparing the least skillful ensemble (the simple average of the original models) and the best performing ensemble (the SVD_KF_ensemble) we find a 33 percent improvement in rmse and 12 percent improvement for the correlation coefficient.

A second way to show the SVD_KF_ensemble improvement is through the spatial distribution of rmse and correlation over the analysis geographical domain (Figure 6). The raw ensemble rmse reaches 20 ppbv mostly around the highly polluted Houston area, and correlation coefficients range from 0.5 to 0.95 across Texas. In contrast, SVD_KF ensemble shows rmse below 12 ppbv and correlation coefficients over 0.75 almost everywhere.

3.3 Categorical statistics.

Categorical statistics are calculated through the creation of the contingency table (Wilks, 1995) with a 10 ppbv increment. Figure 7 shows five statistics: Proportion Correct (PC), Critical Success Index (CSI), Probability of Detection (POD), False Alarm Ratio (FAR), and Frequency Bias (FB). PC is the summation of the diagonal elements of the contingency table divided by the total number of events, so it credits "yes" and "no" correct forecasts equally. Multiplying by 100% it is referred to as the
Percent Correct. CSI is the measure of “yes” forecasts divided by the number of cases forecast and/or observed, so it’s a hit rate for the quantity being forecast, after removing correct “no” from consideration. POD or Hit Rate is the forecast measure of “yes” forecast over all observed events. FAR is the proportion of forecast events that fail to materialize. And FB is the number of forecasts divided by the number of observations within each bin. For the perfect forecast all of these (except FAR) should be unity, and FAR should be zero. In most cases the raw ensemble is at best only marginally superior to persistence. In contrast, the SVD_KF ensemble out performs the raw ensemble and persistence for almost all categories and all O_3 bins. Thus, for example, Percent Correct reaches 50% for SVD_KF ensemble compared to 30% of the raw ensemble and even less for persistence. The only category that does not show better performance over persistence is the Frequency Bias. However, the SVD_KF ensemble is still much closer to unity than the raw ensemble. We note that because the raw ensemble has no ozone values greater than 90 ppbv, that no data points are plotted in Fig. 7 at the two highest ozone bins for the raw ensemble for FAR, and that CSI, POD, and FB are zero.

It is also important to note the improvement of the Heidke skill score, which is the percentage improvement in forecast accuracy compared to random chance. For the raw ensemble, HSS is 17%, but for SVD_KF_ensemble it is more than twice as large, 36%.

3.4 Probability statistics.

One of the significant advantages of ensembles is that they provide probabilistic information and a quantitative estimate forecast uncertainty. From the variety of methods used in the meteorological community, we choose three: Rank Histogram (RH), Attributes Diagram (AD) and Receiver Operating Characteristic (ROC) (Wilks, [1995]).

Rank Histogram (Hamill, [2001]) is calculated by constructing a vector of the 7 models and the
corresponding observation value for the same day and site into an 8-point vector which is then
sorted from the highest to the lowest $O_3$ value. The position of the observation value (or its rank) in
the vector is calculated. The same process is repeated all over possible sites and forecast times. It
provides a quick examination of some qualities of the ensemble. Thus, consistent biases in the
ensemble forecast will show up as a sloped histogram; a lack of variability in the ensemble will show
up as a U-shaped histogram; and in a perfect ensemble the observed value will equally occur in each
rank.

In Figure 8(a), the rank histogram clearly shows an existing bias of the raw model ensemble
forecast, no bias but too small variability in the 7DRM_ensemble forecast, and a better and more
uniform rank histogram for the KF_ensemble.

In Figures 8(b) (for AD) and 8(c) (for ROC) ensemble statistics are presented for an $O_3$ threshold of
70 ppbv. On the Attribute Diagram, which is simply a plot of forecast probability versus the observed
frequency of event, all three ensembles show skilled behavior but the KF_ensemble is closer to the
ideal diagonal line, and almost perfect up to a 60% forecast probability.

The ROC diagram is a plot of Hit Rate versus False Alarm Rate, which is calculated over a simple
2x2 contingency table of the seven models probabilistic forecast (ensembles) stratified into 1/7 wide
categories. For example, for the 4/7 category the Hit Rate is the number of events where 4 or more
models predict $O_3 > 70$ ppbv, divided by the number of occurrences, and False Alarm Rate is the
number of events when 4 and more models predict $O_3 < 70$ ppbv, divided by the number of non-
occurrences. For a perfect forecast, the Hit Rate is 1 and False Alarm Rate is 0, which is the left upper
corner of the plot. Figure 8(c) shows that the KF_ensemble significantly outperforms both the raw
and 7DRM ensembles.

4. PM$_{2.5}$ comparison.
PM$_{2.5}$ has much less diurnal variation compared to O$_3$, and the EPA standard for PM$_{2.5}$ is the daily averaged value. Following this criterion, we plot the time series of the observed and 7 model values of daily PM$_{2.5}$ for the 38 PM stations in the AIRNOW network (Figure 9). From this figure a large discrepancy between the observed and modeled values occurred for most (but not all) sites on August 27-30, 2006, days 14-17. Through evaluation of satellite images, this discrepancy was found to be related to a “Sahara dust event” that brought dust from Africa into the Gulf of Mexico. Eventually, this dust advected as far north as 31 deg latitude, before northerly winds brought in cleaner air. To eliminate the Sahara dust influence in the data, we omitted PM$_{2.5}$ values between August 27-30, 2006, as indicated by the black vertical bars in Fig. 9, for all sites south of 31 degrees latitude.

Using the same techniques that were applied to O$_3$, we generated the ensemble time series of hourly PM$_{2.5}$ values: raw ensemble, 7DRM ensemble, KF ensemble, SVD ensemble, SVD_7DRM ensemble, and SVD_KF ensemble. For the first comparison, we calculated the diurnal cycle of the ensembles (Figure 10). In contrast to O$_3$, PM$_{2.5}$ has a double peak during the day. Even the raw ensemble, which has no bias correction, follows this pattern, as did all of the individual models.

Figure 11 displays PM$_{2.5}$ bulk statistics parameters, including bias, root mean square error and correlation coefficient. Again, the solid bars represent the original models, the hatched bars represent 7DRM bias corrected models, and open bars represent KF models. Also, the solid line is a climatology value and the dashed line indicates a persistence forecast value. These diagrams indicate a significant improvement of the bias corrected models (both 7DRM and especially KF) over the raw models, and further improvements in the SVD raw and KF ensembles. We note that not a single raw or 7DRM bias corrected model is able to do better than persistence, for either rmse or correlation, and that only a few individual KF models have better performance than persistence.

At the far right end of Figure 11, there are all six types of ensembles. Each ensemble performs
better than the individual models that form the ensembles (except for the BAMS correlation
coefficient compared to the raw ensemble), but only the KF ensemble and SVD_KF ensemble are
capable of significantly beating the persistence forecast. The overall improvement of the SVD_KF
ensemble compared to the raw ensemble is 43% for rmse and more than 62% for the correlation
coefficients. All ensemble statistics comparisons are shown in Table 2.

In terms of categorical statistics, the raw ensemble does not show much improvement compared
to the persistence forecast, but the SVD_KF ensemble does (Figure 12). Even for the Frequency Bias,
the SVD_KF ensemble has a value near unity through a wide range up to 23 µm/m³. The raw
ensemble Heidke skill score is 14%, while the SVD_KF ensemble score equals 34%.

Probability statistics, in terms of Rank Histogram, Attribute Diagram and Receiver Operating
Characteristic (ROC), all show much better performance for both bias-corrected ensembles
compared to the raw ensemble (Figure 13). The slope of the Rank Histogram for the raw ensemble
indicates a bias, while the KF model’s Rank Histogram has the most uniform shape, Figure 13(a).
The Attribute Diagram and ROC are presented for a PM$_{2.5}$ threshold of 20 µm/m³. In the Attribute
Diagram, Figure 13(b), the 7DRM and KF ensembles most often have values closer to the diagonal
ideal, and in the ROC curve, Figure 13(c), the KF ensemble points are the closest to the left upper
corner (high hit rate and low false alarm rate).

5. Conclusions.

In this study several ensembles were formed from seven air quality forecast models and are
evaluated using O$_3$ and PM$_{2.5}$ observations from the EPA’s AIRNOW surface network. The
observations were collected and all models were running in real time during the TexAQS II
experiment in late summer of 2006. From the original models two bias-corrected sets of models
were developed, using either a running mean or a Kalman filter approach, utilizing the 7 previous
days of data.

Six different types of ensemble were generated from the raw and two bias-corrected sets of
model forecasts. First were the raw ensemble, 7DRM ensemble and KF ensemble. Second, weight
coefficients were calculated (again using the previous 7 days of data) for each hour of the day which
were used in the weighted averaging of three sets of models. Thus, three other ensembles were
developed: SVD ensemble, SVD_7DRM ensemble and SVD_KF ensemble.

Bulk statistics show that each ensemble is almost always as good or better than the best of the
individual models, the Kalman Filter models have improved statistics compared to 7DRM models, and
the SVD method adds further improvements to the corresponding ensembles. The combination of KF
bias correction and a multi-model weighted averaging provides the highest predictive skill. Thus, the
SVD_KF ensemble of O₃ shows a 33% improvement compared to the raw ensemble for rmse and a
12% improvement in the value of the correlation coefficients. For PM₂.₅, these gains are even more
impressive: a 43% improvement for root mean square error and a 62% gain for the mean correlation
coefficient. The large improvement probably comes from the fact that each of the individual models
poorly predicted surface PM₂.₅ during the TEXAQS-II experiment. All of the original models, the 7DRM
bias-corrected models and most of KF models underperform compared to a persistence forecast of
PM₂.₅. Only a few KF models and the SVD_KF ensemble are significantly better than persistence.

Except for Frequency Bias, categorical statistics almost always show improved skill for the SVD_KF
ensemble compared to persistence, which is not always the case for the raw ensemble. By nature,
the ensemble forecast is a probabilistic forecast. Several probabilistic measures, constructed for the
different types of ensembles, demonstrate that the Kalman Filter bias-corrected models are better
than the raw and 7DRM corrected forecasts. This includes a more uniform Rank Histogram, more
accurate probability forecasts shown in the Attribute Diagram, and higher Hit Rates/lower False
Alarm Rates in the ROC diagram.

The above results confirm the advantage of the bias correction methods (7DRM and KF) for air
quality forecasting. Both of these methods require data only from the 7 previous days, therefore the
bias-corrected techniques may easily be applied to any operational forecast. Further improvement
can be reached using the ensemble of the corrected models, either evenly or as a weighted average.
In particular we note that the skill improvements from the bias correction and ensemble techniques
are greater for a variable with low-skill forecast (PM$_{2.5}$) than for ozone.

Acknowledgements.
This research is partially funded by Early Start Funding from the NOAA/NWS Office of Science and
Technology, and the NOAA Office of Oceanic and Atmospheric Research/Weather and Air Quality
Program. Credit for program support and management is given to Paula Davidson (NOAA/NWS/OST)
and Jim Meagher (NOAA/ESRL/CSD).
References:


Otte, T. L., et al., 2005, Linking the Eta model with the Community Multiscale Air Quality (CMAQ) modeling system to build a national air quality forecasting system, Weather and Forecasting, 20, 367-384.


Parrish, D. D., et al. (2009), Overview of the Second Texas Air Quality Study (TexAQS II) and the Gulf of Mexico Atmospheric Composition and Climate Study (GoMACCS), J. Geophys. Res., 114, D00F13, doi:10.1029/2009JD011842.


van Loon, M., at al., Evaluation of long-term ozone simulations from seven regional air quality models and their ensemble, 2007, Atmos. Environ., 41, 10, 2083-2097

Vautard, R., at al., 2007, Evaluation and intercomparison of Ozone and PM10 simulations by several chemistry transport models over four European cities within the CityDelta project, J. Atmos. Environ., 41, 1, 173-188

Vautard, R., at al., 2008, Skill and uncertainty of a regional air quality model ensemble, Atmos. Environ., 43, 31, 4822-4832


Figure Captions:

Figure 1. Spatial distribution of the O$_3$ and PM$_{2.5}$ AIRNOW network surface observation sites during Texas Air Quality Study, 2006, field study.

Figure 2. Composite diurnal cycle of surface O$_3$ concentrations, from the observations (black) and 7 models (colored), over the 48-hour forecast cycle, averaged over 50 days at 118 sites in Texas during the summer of 2006. Also shown by the red box is the time period during which the daily values of 8-hour maximum O$_3$ were calculated.

Figure 3. Mean O$_3$ over the 48-hour forecast cycle averaged over all sites and all days of the three ensembles and the observations (top panel). Ensemble biases are shown at the bottom panel.

Figure 4. Time series of 8-hour maximum O$_3$ from the observations (black) and models (colors) for the original (a), 7-day running mean bias-corrected (b), and 7-day Kalman-Filter-corrected (c) model data, following the color scheme for the various models listed at the top of the figure. The averaged O$_3$ time series for all sites is shown in red for the observations and in black for the models. Thin vertical black lines are plotted for each Saturday.

Figure 5. O$_3$ bias (top panel), rmse (middle panel) and correlation coefficients (bottom panel) for the 7 models. The skill from climatology is shown as the solid line, and the value for persistence is shown as the dashed line. The solid colored boxes show the original models, the hatched boxes show the 7DRM bias-corrected models, and the open boxes show the KF bias-corrected models. Also shown are the skills of the three ensembles, calculated from the three various sets of the models: the raw ensemble (blue), the 7DRM ensemble (hatched red) and the KF ensemble (open green), and their SVD modifications. The improvement of rmse and correlation coefficients performance of the SVD_KF ensemble compared to the raw ensemble is marked.

Figure 6. Spatial distribution of the raw ensemble and the SVD_KF_ensemble O$_3$ root mean square error (RMSE) and correlation coefficient (CORR) over Texas region.
Figure 7. Categorical statistics of 8-hour maximum O₃: PC-proportion correct, CSI-critical success index, FAR-false alarm rate, POD-probability of detection and FB-frequency bias. In all categories the SVD_KF_ensemble statistics are shown in the thick solid lines, the raw ensemble statistics are shown as the thin solid lines and the value of persistence by the dashed lines.

Figure 8. Probabilistic statistics for O₃ including the rank histogram (a), attribute diagram for O₃ > 70 ppbv (b) and receiver operating characteristic (ROC) for O₃ > 70 ppbv (c). The raw ensemble is shown in light lines, the 7DRM ensemble is shown in medium lines, and the KF ensemble is shown in dark lines.

Figure 9. Time-series of 24-hour averaged PM₂.₅ for all 38 sites from the observations (top panel) and the 7 models, following the color scheme for the various models listed at the top of the figure. Overlaid on the times series are the site-averaged PM₂.₅, shown in red for the observation data and in black for the models. The black box indicates Sahara dust event that occurred between August 27-30, 2006.

Figure 10. Mean PM₂.₅ over the 48-hour forecast cycle averaged over all sites and all days for the six ensembles and the observations (top panel), following the color scheme at the top of the figure. Ensemble biases are shown at the bottom panel.

Figure 11. As in Figure 5 except for PM₂.₅ bulk statistics.

Figure 12. As in Fig. 7 except for PM₂.₅ categorical statistics.

Figure 13. As in Figure 8 except for PM₂.₅, with a threshold of 20 μm/m³ for the Attribute Diagram and Receiver Operating Characteristic.

Figure 8 (color). Probabilistic statistics for O₃ including the rank histogram (a), attribute diagram for O₃ > 70 ppbv (b) and receiver operating characteristic (ROC) for O₃ > 70 ppbv (c). The raw ensemble is shown in blue, the 7DRM ensemble is shown in red, and the KF ensemble is shown in green.
Fig. 1.
Fig. 2

Fig. 3
Fig. 5

Fig. 6
CATEGORICAL FORECAST STATISTICS

Fig. 7

CATEGORICAL FORECAST STATISTICS

Fig. 7 (WE)
Fig. 8
Fig. 8 (WB)
Fig. 11
Fig. 13
Fig. 13 (WB)