Improving Air Quality Predictions over the United States with an Analog Ensemble

LUCA DELLE MONACHE[a] AND STEFANO ALESSANDRINI

National Center for Atmospheric Research, Boulder, Colorado

IRINA DJALALOVA AND JAMES WILCZAK

National Oceanic and Atmospheric Administration, Boulder, Colorado

JASON C. KNIEVEL AND R. KUMAR

National Center for Atmospheric Research, Boulder, Colorado

(Manuscript received 29 July 2019, in final form 2 March 2020)

ABSTRACT

Air quality forecasts produced by the National Air Quality Forecasting Capability (NAQFC) help air quality forecasters across the United States in making informed decisions to protect public health from acute air pollution episodes. However, errors in air quality forecasts limit their value in the decision-making process. This study aims to enhance the accuracy of NAQFC air quality forecasts and reliably quantify their uncertainties using a statistical–dynamical method called the analog ensemble (AnEn), which has previously been found to efficiently generate probabilistic forecasts for other applications. AnEn estimates of the probability of the true state of a predictand are based on a current deterministic numerical prediction and an archive of prior analogous predictions paired with prior observations. The method avoids the complexity and real-time computational expense of model-based ensembles and is proposed here for the first time for air quality forecasting. AnEn is applied with forecasts from the Community Multiscale Air Quality (CMAQ) model. Relative to CMAQ raw forecasts, deterministic forecasts of surface ozone (O_3) and particulate matter of aerodynamic diameter smaller than 2.5 μm ($PM_{2.5}$) based on AnEn's mean have lower systemic and random errors and are overall better correlated with observations; for example, when computed across all sites and lead times, AnEn's root-mean-square error is lower than CMAQ's by roughly 35% and 30% for O_3 and $PM_{2.5}$, respectively, and AnEn improves the correlation by 50% for O_3 and $PM_{2.5}$. Probabilistic forecasts from AnEn are statistically consistent, reliable, and sharp, and they quantify the uncertainty of the underlying prediction.

1. Introduction

Every year poor air quality kills millions of people worldwide (Forouzanfar et al. 2015), and in the United States alone it costs society from tens to hundreds of billions of dollars (Muller and Mendelsohn 2007). Air quality forecasts are one resource that decision-makers can use to reduce many threats that poor air quality poses. However, uncertainty in predictions can reduce their value in the decision-making process.

A reliable quantification of uncertainty in air quality forecasts is crucial for determining their value in the decision-making process. Unfortunately, uncertainty cannot be completely eliminated from air quality forecasting, but there are effective ways to treat the inevitable uncertainty and to reduce its consequences. One way is through probabilistic ensemble prediction. In contrast to the *deterministic* approach of using a single forecast from a single model, *probabilistic* information is obtained from an ensemble that comprises multiple and meaningfully different forecasts that are valid at the same future time and location.

Ensembles are beneficial in many ways. The probabilistic guidance they provide is potentially much more useful for decision-makers than a single forecast could ever be (Buizza 2008; Palmer 2002). An ensemble's

[a] Current affiliation: University of California, San Diego, La Jolla, California.

Corresponding author: Dr. Luca Delle Monache, ldellemonache@ucsd.edu

DOI: 10.1175/WAF-D-19-0148.1

mean forecast tends to be (but is not always) more skillful than any individual member's prediction (Delle Monache et al. 2006a,b, 2008; Delle Monache and Stull 2003; Delle Monache 2010; Djalalova et al. 2010; Du et al. 1997; Ebert 2001; Galmarini et al. 2001, 2004; Kioutsioukis et al. 2016; Leith 1974; McKeen et al. 2005; Potempski et al. 2008; Potempski and Galmarini 2009; Solazzo et al. 2012; Toth and Kalnay 1997; Ma et al. 2012). Calculating the mean filters out some of the unpredictable elements of the physical and chemical processes being simulated. Another benefit is that the approximate uncertainty in a mean forecast can be inferred from the spread among ensemble members (Kalnay 2003) if the ensemble is calibrated, although that inference has to be made carefully and is not valid in every case (Barker 1991; Hopson 2014; Murphy 1988). Ensembles also produce *reliable* and *well resolved* probabilistic forecasts that can be further improved through calibration and other methods of postprocessing imperfect numerical predictions. A *reliable* ensemble is one that over many cases predicts conditions to occur with the same frequency as they actually occur in nature. A *well-resolved* ensemble is one that provides a probability close to 100% on occasions when an event (e.g., ozone above 100 ppb) occurs and forecast close to 0% when the event does not occur (i.e., it is specific from case to case about whether or not a condition will occur).

In air quality forecasting in particular, probabilistic approaches have been recommended as—and have been demonstrated to be—effective at dealing with the many sources of uncertainty (e.g., Bei et al. 2010; Carmichael et al. 2008; Dabberdt et al. 2004; Delle Monache et al. 2006a,b, 2008; Delle Monache and Stull 2003; Garaud and Mallet 2010; Kioutsioukis et al. 2016; Marécal et al. 2015; Mallet 2010; Mallet et al. 2013; Mallet and Sportisse 2006a,b; McKeen et al. 2005; Pagowski et al. 2005; Zhang et al. 2007, 2012). Uncertainties stem from meteorological initial and boundary conditions; the sparse temporal and spatial distribution of observations, and their errors, which are assimilated into a model; truncations and approximations in a model's numerical schemes; uncertainties in emission inventories, which are often not well known nor well characterized in models; and physical or chemical processes that are simplified, poorly understood, or omitted entirely (Delle Monache and Stull 2003).

Ensemble prediction takes different forms. One of the simplest is a lagged ensemble (Dalcher et al. 1988; Delle Monache et al. 2006a; Ebisuzaki and Kalnay 1991; Hoffman and Kalnay 1983; Lu et al. 2007; Mittermaier 2007). In a lagged ensemble, sequential forecasts from a deterministic system are grouped together, each with a common valid time but with different lead times (e.g., an 18-h forecast initialized at 0000 UTC and valid at 1800 UTC, a 15-h forecast initialized at 0300 UTC and valid at 1800 UTC, and so on). The effectiveness of this approach depends on how frequently the deterministic forecasts are updated (Lu et al. 2007; Mittermaier 2007).

Lagged ensembles target uncertainty in the initial conditions. Similarly, single-model ensembles use different initial conditions, boundary conditions, and/or perturbed observations to generate diversity even though the model configuration (including physical parameterizations) is fixed (e.g., Molteni et al. 1996; Zhang et al. 2007). In modeling air quality, perturbations can be applied not just to meteorological observations, but also to emissions, chemistry, and deposition. Considering different types of perturbations is important for addressing the nonlinearity of the air quality forecasting (Delle Monache et al. 2006a,c).

Ensembles based on multiple models are often more effective but are more complex and computationally expensive. Multimodel ensembles address uncertainty in the actual physical or chemical processes being simulated, not merely in the initial meteorological or chemical conditions (e.g., Bei et al. 2010; Delle Monache et al. 2008; Delle Monache and Stull 2003; Djalalova et al. 2010; Garaud and Mallet 2010; Kioutsioukis et al. 2016; Mallet and Sportisse 2006a,b; McKeen et al. 2005; Vautard et al. 2012). For the purposes of this article, we consider what are sometimes called *multiphysics* ensembles to be just one type of *multimodel* ensemble. Multimodel ensembles have proved to be a particularly effective tool for probabilistic operational weather forecasting for quite some time (e.g., Buizza et al. 2005; Hacker et al. 2011; Krishnamurti 1999). In air quality forecasting, multimodel ensembles have been successfully applied to forecasts of ground-level ozone (e.g., McKeen et al. 2005; Solazzo et al. 2012; Žabkar et al. 2013), airborne particles (e.g., Djalalova et al. 2010; McKeen et al. 2007), and to both (e.g., Monteiro et al. 2013).

The complexity and expense of multimodel ensembles and the limitations of single-model ensembles are two motivations for developing simpler, more cost-effective, yet still powerful methods for probabilistic prediction—especially hybrid statistical–dynamical methods. One of such methods, is the analog ensemble (AnEn), which has been applied successfully to weather parameters (Delle Monache et al. 2013; Sperati et al. 2017; Alessandrini et al. 2018, 2019), and renewable energy (Alessandrini et al. 2015a,b; Junk et al. 2015; Davò et al. 2016; Cervone et al. 2017). The goal of this article is to explore the applicability of AnEn to air quality forecasting. While three different AnEn versions, two of which in combination with a Kalman filter bias

correction, have been proposed for deterministic air quality predictions (Djalalova et al. 2015; Huang et al. 2017), in this paper AnEn is evaluated for the first time for probabilistic air quality predictions. We compare AnEn's forecasts of ground-level ozone (O_3) and particulate matter of aerodynamic diameter smaller than 2.5 μm ($PM_{2.5}$) to the simpler probabilistic standard of a persistence ensemble (PeEn), and to an operational standard of the industry, the U.S. Environmental Protection Agency (EPA) Community Multiscale Air Quality (CMAQ) model (Byun and Schere 2006). The PeEn can be thought as the probabilistic equivalent of a deterministic prediction based on persistence, which can be used as a baseline method for probabilistic predictions when more advanced ensemble systems are not available for comparison purposes, as in this study for air quality predictions. The two ensemble methods, the CMAQ configuration, and the observations used for this research are described in section 2. The performance of the ensemble methods and CMAQ against the observations is assessed in section 3, and the results are summarized in section 4.

2. Predictive systems and data

a. Analog ensemble

AnEn is a hybrid statistical–dynamical method that generates an ensemble for estimating the probability of some future observation of a predictand (e.g., 2-m dewpoint, geopotential height at 500 hPa, or, in the case of air quality forecasts, $PM_{2.5}$) given a current deterministic prediction and an archive of historical, analogous deterministic predictions paired with historical observations at those predictions' valid times. That archive is used to train the AnEn. Typically, the deterministic predictions in the training data and the current deterministic prediction come from the same configuration of the same NWP or air quality model, or nearly so. Hamill and Whitaker (2006) demonstrated the considerable value of using an analog ensemble approach to calibrate an existing ensemble, and Delle Monache et al. (2013) proposed a similar approach to instead generate an ensemble, which forms the basis of our approach.

If the future observation of a predictand is represented by y, then the probability distribution of that future observation is

$$f(y|\mathbf{x}^f), \tag{1}$$

where $\mathbf{x}^f = (x_1^f, x_2^f, x_3^f, \ldots, x_k^f)$ are the k predictors from the deterministic model forecast. Several steps are involved in generating $f(y|\mathbf{x}^f)$ (i.e., how an analog is defined, how its quality is assessed, how many analogs to select from the archive, and how deep an archive is necessary, etc.) and can vary from one application of the AnEn to another. Section 3 goes into more detail about our application's sensitivity to some of those choices.

For determining which historical forecasts are sufficiently analogous to the current forecast, we follow Delle Monache et al. (2011, 2013), Nagarajan et al. (2015), and Djalalova et al. (2015), in using the metric developed by Delle Monache et al. (2011):

$$\|F_t, A_{t'}\| = \sum_{i=1}^{N_v} \frac{w_i}{\sigma_{f_i}} \sqrt{\sum_{j=-\tilde{t}}^{\tilde{t}} (F_{i,t+j} - A_{i,t'+j})^2}, \tag{2}$$

where F_t is the current deterministic forecast for a location of interest, valid at some time t in the future; $A_{t'}$ is an analogous forecast from the archive, valid at some time t' in the past, and with the same lead time as the current forecast's; N_v is the number of atmospheric variables (i.e., predictors) used to select analogs; w_i is the weight assigned to each atmospheric variable of index i; and σ_{f_i} is the standard deviation of historical forecasts of each atmospheric variable of index i. The metric is calculated over a range of times from $-\tilde{t}$ to $+\tilde{t}$ centered on the valid time t. The $F_{i,t+j}$ and $A_{i,t'+j}$ respectively are each current forecast and each analogous historical forecast for atmospheric variable index i within that range of times. We set $\tilde{t} = 1$ h. The forecast interval is 0–48 h.

The weights w_i for each analog predictor were determined independently for each observing site as explained in section 3b. From each search of the archived datasets, the best 20 analogs were chosen, which for any given search is 3%–4% of the total cases in the archive (section 2c). The selection of the number of analogs to be used is based upon a balance between sampling enough of the observed distribution of the predictand variable while ensuring that all analogs are sufficiently similar to the current prediction.

b. Persistence ensemble

Like Alessandrini et al. (2015a), we use a persistence ensemble as a baseline method for probabilistic prediction and then demonstrate AnEn's improvement on that baseline. For each forecast lead time, PeEn is based on the most recent 20 observations of O_3 or $PM_{2.5}$ at the same hour of day as the forecast valid time. Other tested PeEn configurations (which do not perform as well as the latter, not shown) include PeEn formed by the most recent observations collected the previous seven days over a 3-h window centered on the

same hour of the day, and a configuration including observations from the previous four days over a 5-h window. The PeEn ensemble can be skillful when similar air quality conditions persist for several days, or when conditions fluctuate with the same repeating diurnal pattern. On the other hand, it is challenging for PeEn to capture rapidly changing patterns of O_3 or $PM_{2.5}$.

c. CMAQ and predictors

Air quality forecasts used in this study are based on the CMAQ model, version 5.02 (Byun and Schere 2006), which is the official Chemistry–Transport Model (CTM) adopted by the National Air Quality Forecasting Capability (NAQFC) for operational air quality predictions over the United States. It is a modular, Eulerian, Cartesian modeling system that simulates the emission, production, advection, diffusion, chemical transformation, and removal of atmospheric pollutants at regional scales. CMAQ's daily forecasts of ground-level concentrations of O_3 and $PM_{2.5}$ at lead times of 0–48 h and horizontal grid spacing of 12 km are provided as input to the AnEn's algorithm. Additional inputs for AnEn include 10-m wind speed and direction, 2-m air temperature, 2-m specific humidity, and cloud cover, which were extracted from the National Centers for Environmental Prediction (NCEP) North America Model output that provides meteorological fields necessary to drive CMAQ operationally.

The rational for selecting the aforementioned air quality and meteorological variables as predictor variables, which we recognize may not be exhaustive, is as follows. O_3 and $PM_{2.5}$ allow us to identify pollution episodes of similar magnitude in the past. Temperature plays a vital role in several processes relevant to air quality including atmospheric chemical kinetics, biogenic emissions, and mixing. The wind speed and wind direction allow us to assure that similar transport pathways contributed to the analogous air pollution episodes in the past. Humidity is selected for its key role in the formation and destruction of both O_3 and $PM_{2.5}$. The water vapor (H_2O) in conjunction with O_3 photolysis is the main source of hydroxyl (OH) radical which in turn initiates photochemical production of O_3 through oxidation of different volatile organic compounds (VOCs). In the case of $PM_{2.5}$, humidity determines the aerosol water content, which is important for secondary aerosol formation. Cloud cover determines the amount of solar radiation available for atmospheric photochemical reactions that produces both O_3 and $PM_{2.5}$. In summary, the predictors are strategically selected in such a way that they are not only able to identify the pollution episodes of similar magnitude in the past but also identify the meteorological and chemical conditions leading to similar air pollution episodes in the past.

d. Observations

The source of observations is the EPA AIRNow network (EPA 2017) in the conterminous United States and southern Canada (Fig. 1). Hourly concentrations of O_3 and $PM_{2.5}$ are obtained from 1337 and 551 sites, respectively. All of the observations used in this study are subjected to a quality control procedure that is suitable for real-time operational forecasting and described in detail for $PM_{2.5}$ in section 2 of Djalalova et al. (2015). Furthermore, observation sites frequently reporting missing data are excluded from the analysis presented in this study (i.e., only stations with at least 50% of data available are retained). This results in 1045 and 458 sites for O_3 and $PM_{2.5}$, respectively, which are then used to generate AnEn.

3. Results

This section begins with examples of AnEn's air quality forecasts, followed by sensitivity tests of the ensemble's algorithm as the number of analogs and the length of the training data are varied, and a description of the analog predictor weights. An in-depth analysis of AnEn's performance compared to CMAQ's (for deterministic predictions) and PeEn's (for probabilistic predictions) is presented afterward. The periods of study for O_3 is from 1 July 2014 to 30 September 2015 (456 days). Given that O_3 is a major air quality problem during summertime, that has been broken down in a training period from 1 July 2014 to 31 May 2015 followed by a verification period from 1 June to 30 September 2015. Similarly, for $PM_{2.5}$ whose concentration is higher in wintertime, the period of study from 1 July 2014 to 29 February 2016 (608 days), has been divided in a period of training from 1 July 2014 to 30 November 2015 and a period of verification from 1 December 2015 to 29 February 2016.

a. Examples of forecasts from AnEn, PeEn, and CMAQ

Figure 2 shows two randomly chosen examples of O_3 (top panels) and $PM_{2.5}$ (bottom panels) predictions by the methods considered in this study. For O_3, both the AnEn and PeEn ensemble means reduce some of the CMAQ biases, particularly at night when CMAQ substantially overestimates O_3 concentration. State-of-the-art CTMs generally struggle to simulate nighttime O_3 because of the challenges in representation of stratified boundary layer. AnEn's mean is closer to the observations than the PeEn's mean. The usefulness of

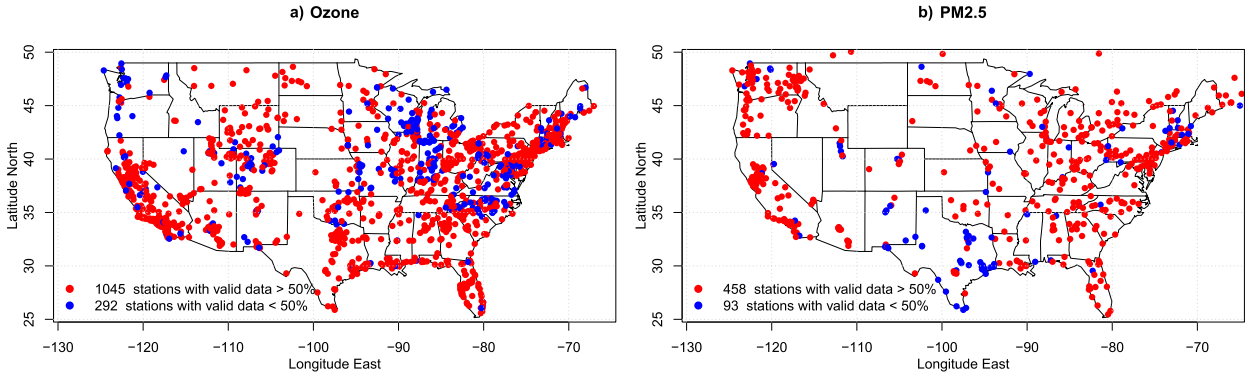

FIG. 1. (a) Ozone and (b) $PM_{2.5}$ observation sites used in this study. Shown are sites for which hourly observations are available for more (red) or less (blue) than 50% of the time during the two study periods for O_3 and $PM_{2.5}$, which are described at the beginning of section 3.

probabilistic predictions is evinced for O_3 at forecast lead times 28–30, when the deterministic predictions by CMAQ and the ensemble means miss the observed peak, but the ensemble spread from both AnEn and PeEn indicates a low probability of higher concentration, which could be useful information for a decision-maker trying to protect the public from pollution episodes. Similar qualitative performance of both the ensembles is also seen for the $PM_{2.5}$ predictions (Fig. 2, bottom panels).

b. *Sensitivity analysis of the analog ensemble*

AnEn's performance is sensitive to the number of analogs chosen from the historical dataset (Fig. 3).

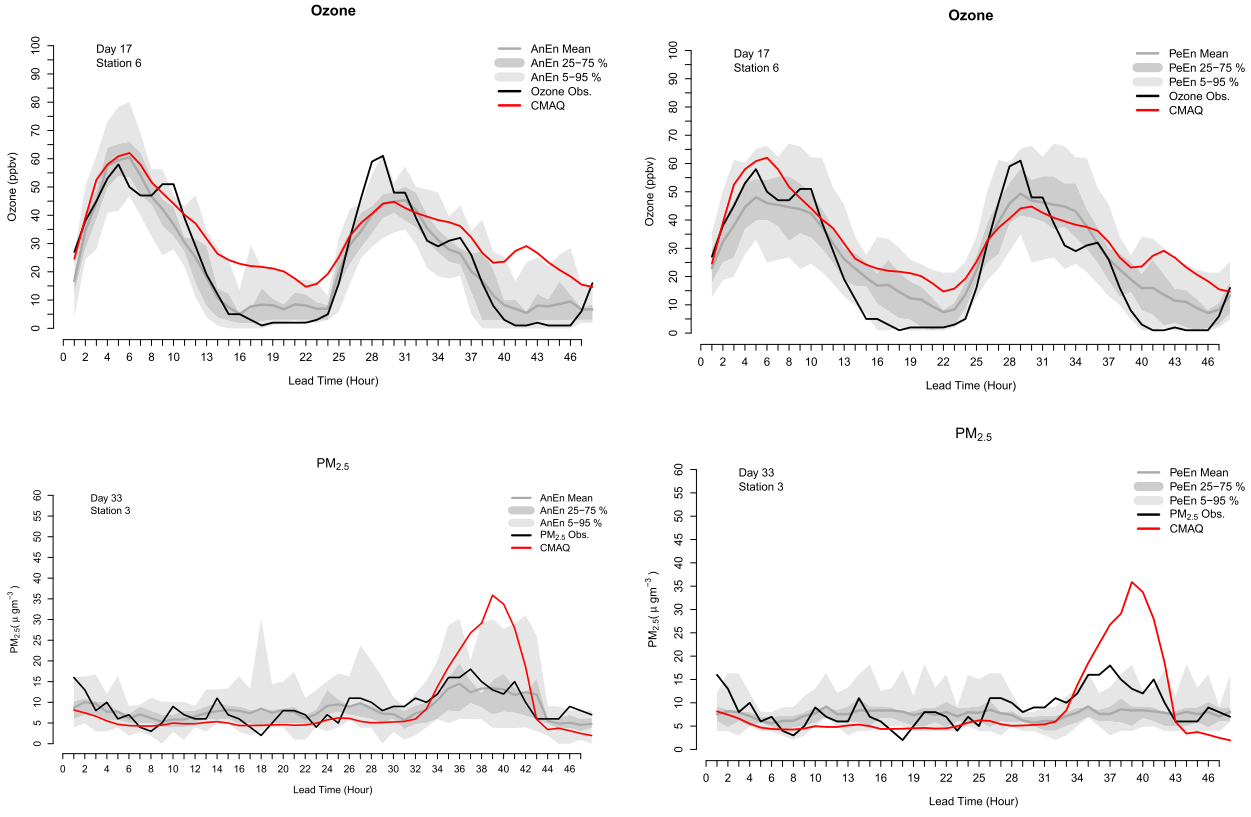

FIG. 2. Randomly chosen examples of 0–48-h predictions of (top) O_3 and (bottom) $PM_{2.5}$ for (left) AnEn and (right) PeEn at two different locations and days. CMAQ predictions are in red, and the observations are in black. Both ensemble distributions are depicted with the mean (dashed line) and the 5–95 and 25–75 interquantile ranges (shading).

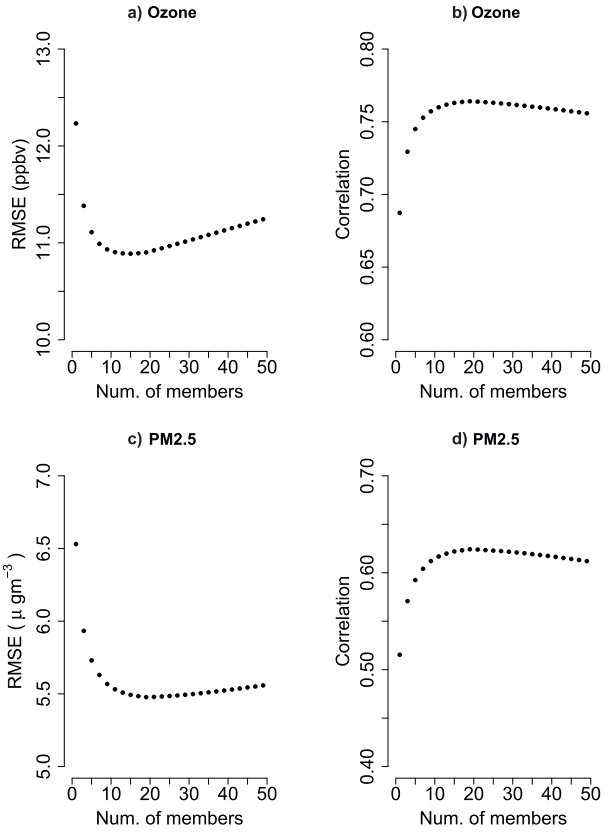

FIG. 3. Sensitivity of AnEn to the number of ensemble members (i.e., analogs). Shown are AnEn forecasts' RMSE of (a) O_3 (ppbv) and (c) $PM_{2.5}$ ($\mu g\,m^{-3}$) and correlation of (b) O_3 (0.0–1.0) and (d) $PM_{2.5}$ (0.0–1.0) vs number of ensemble members (i.e., analogous forecasts selected from the training archive). Calculations are averages over all lead times and sites during the periods of study described in the text. Note the different ranges among the *y* axes.

In this study, 15–25 members produced ensemble mean forecasts of O_3 and $PM_{2.5}$ with the highest correlations and forecasts of $PM_{2.5}$ with the lowest RMSEs. The lowest RMSEs from O_3 forecasts were achieved with 10–20 members. This sensitivity motivated our choice of 20 members for this particular study. Other studies will exhibit different sensitivity and might call for different measures of AnEn's performance, not simply RMSE and correlation.

Figure 3 depicts variations in correlation and RMSE as a function of the number of analog members for both ozone and $PM_{2.5}$. The curves in correlation and RMSE result from two opposing trends. The increase in the number of analogs leads to a more thorough sampling of the statistical relationships between forecasts and observations in the training data. However, inclusion of each additional analog decreases the similarity between it and the current forecast. Under typical circumstances, the shorter the training period the less likely it is to find closing matching analogs, which will normally lead to lower correlations and higher RMSEs.

Another way to measure this sensitivity is to evaluate AnEn's performance as a function of training period with the number of analogs held constant (Fig. 4). Longer training periods improve AnEn's performance. The degree of improvement depends on the variable and metric. Forecasts of O_3 are improved the most. For $PM_{2.5}$, there are very small improvements with longer trainings, particularly for RMSE, with the differences across the experiments being not statistical significant. This may reflect the fact that CMAQ is better correlated to the observations for ozone than for $PM_{2.5}$, which may facilitate the algorithm in finding better analogs with a longer training when applied to ozone.

The weights w_i for each of the five variables are determined independently for each observing site according to an algorithm that minimized continuous ranked probability score (CRPS) over the optimization periods of May 2015 (O_3) and November 2015 ($PM_{2.5}$). The details on analog predictor optimization strategies can be found in Junk et al. (2015). The optimization periods do not overlap with the period over which the performance metrics have been calculated. Weights can have values in the 0.0–1.0 range with increments of 0.1, and the weights of CMAQ O_3 and $PM_{2.5}$ predictors are set to a minimum of 0.4 for the prediction of O_3 and $PM_{2.5}$, respectively. Figure 5 shows the distribution of the weights for each predictor for both O_3 (Fig. 5a) and $PM_{2.5}$ (Fig. 5b). For the prediction of O_3, the predictors are wind speed (WSPD; m s^{-1}) and direction (WDIR; degrees from N) at 10 m AGL, air temperature (T2M;, °C) at 2 m AGL, cloud fraction (CF), and ground-level concentrations of O_3 (ppbv). For $PM_{2.5}$, the analog predictors are T2M, WSPD, WDIR, specific humidity Q ($g\,kg^{-1}$), and surface $PM_{2.5}$ ($\mu g\ m^{-3}$).

The distributions show the weights' variability across the stations. Both O_3 and $PM_{2.5}$ are weighted high for their respective predictions, as expected. For O_3, T2M, WSPD, and WDIR are weighted similarly, while the distribution of the weights for CF corresponds to the lowest values, possibly reflecting the model lower skill in predicting this variable. For $PM_{2.5}$, T2M has the second-highest median of its weight distribution followed by WSPD, WDIR, and then Q.

c. *Deterministic predictions*

When deterministic predictions are evaluated over all forecasts and observations in our study, AnEn's mean forecasts of O_3 and $PM_{2.5}$ are dramatically superior to the raw forecasts from CMAQ (Fig. 6).

The AnEn significantly improves CMAQ's raw prediction by reducing RMSE by approximately 35% (O_3)

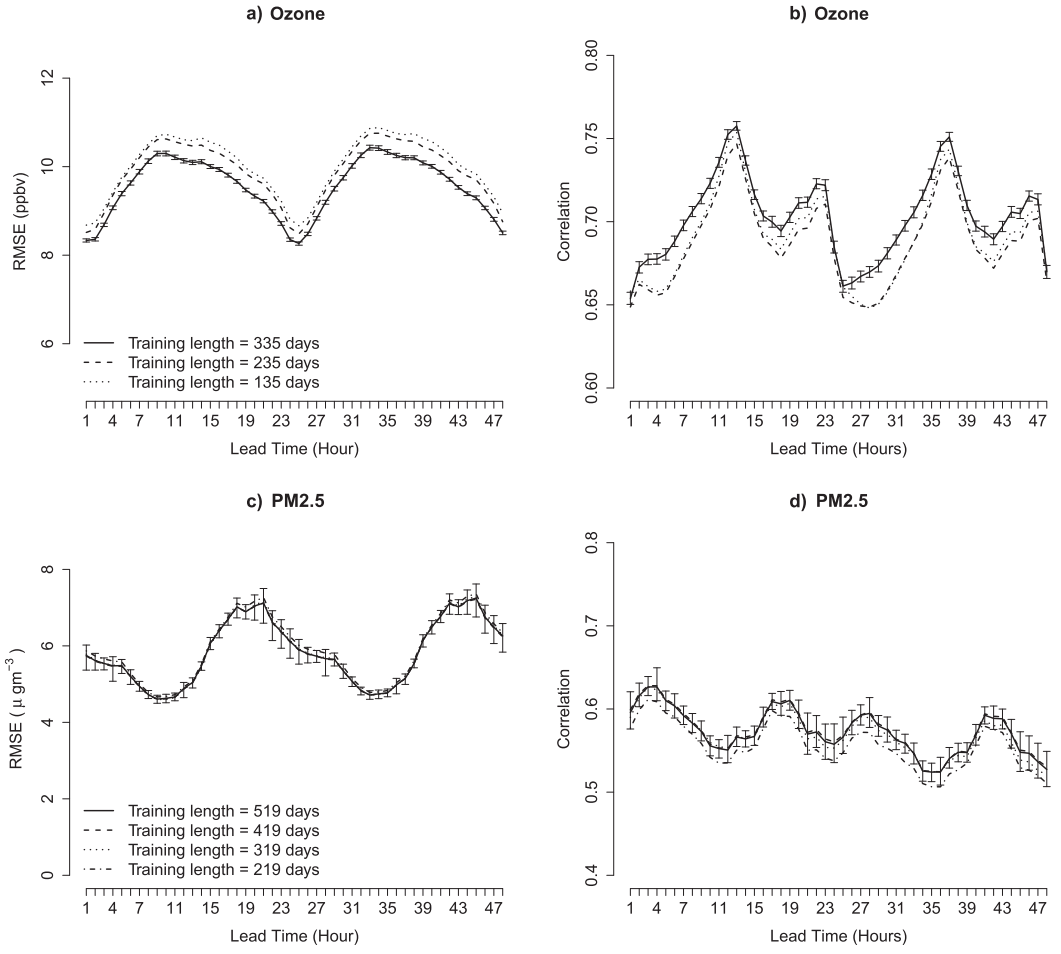

FIG. 4. Sensitivity of AnEn to length of the training period. Shown are AnEn forecasts' RMSE of (a) O_3 (ppbv) and (c) $PM_{2.5}$ (μg m^{-3}) and correlation of (b) O_3 (0.0–1.0) and (d) $PM_{2.5}$ (0.0–1.0) vs lead time (h) for the training periods indicated by the line styles. Calculations are averages over all sites during the periods of study described in the text. The number of analogs is 20 in every case. The vertical bars indicate the 95% confidence intervals computed with bootstrapping, which is applied only to the longest training period to reduce clutter.

and 30% ($PM_{2.5}$) and bias by 90% (O_3) and 95% ($PM_{2.5}$), and by improving the correlation by 50% (O_3 and $PM_{2.5}$) when these performance metrics are computed across all sites and lead times. The metrics in Fig. 6 vary with lead time because there is diurnal variation in CMAQ's skill—which in turn affects AnEn's performance—and because observations are distributed across several time zones.

The AnEn significantly improves CMAQ estimates, which have been used to generate it. As shown in previous contributions (e.g., Delle Monache et al. 2011; Djalalova et al. 2015; Huang et al. 2017), it reduces both systematic and unsystematic errors, while significantly improving the correlation with observations. In principal, given that the analog ensemble estimates are based on past observations, the AnEn mean should provide unbiased estimates. However, the residual bias after the AnEn correction (Figs. 6c,d), is likely due by the fact that the training dataset is finite, which does not guarantee that the distribution of the observations that are the foundation for AnEn fully sample the predicted PDF.

Figures 7 and 8 show the spatial distribution of the AnEn improvements (%) over CMAQ in RMSE and correlation for both O_3 and $PM_{2.5}$, computed independently at each available observation site and over the verification period. For O_3, AnEn improves upon CMAQ (Fig. 7, top panel) almost at every location, with RMSE reductions up to 60%, and similar improvements to correlation (Fig. 8, top panel), although not as pronounced as to RMSE. AnEn improves CMAQ across different land uses, topographical complexities, and geographic regions, resulting in a capability that can be considered for real-time forecasting in operational

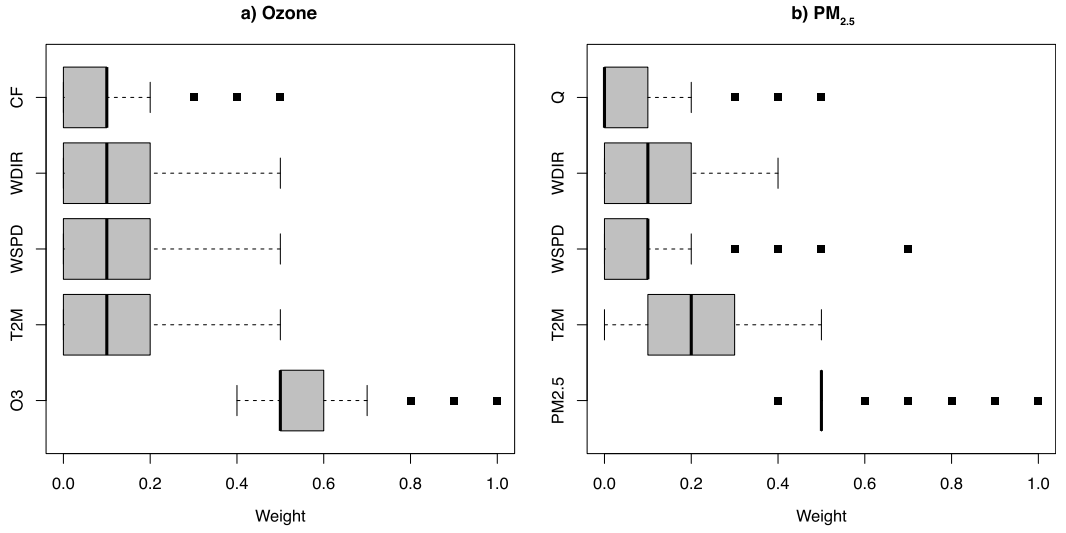

FIG. 5. Distribution of analog predictor weights across the available stations for (a) O_3 and (b) $PM_{2.5}$. The gray boxes indicate the 25–75 interquantile range, the black line within the box is the median, filled squares are the outliers, and the vertical black lines at the edge of the dashed lines are the minimum and maximum excluding the outliers. Ozone predictors include wind speed (WSPD; m s^{-1}) and direction (WDIR; degrees from north) at 10 m AGL, air temperature (T2M; °C) at 2 m AGL, cloud fraction (CF), and ground-level concentrations of O_3 (ppbv). The $PM_{2.5}$ predictors used in this study are T2M, WSPD, WDIR, specific humidity (Q; g kg^{-1}), and surface $PM_{2.5}$ (μg m^{-3}).

centers. For $PM_{2.5}$, several stations indicate instead a degradation after the AnEn correction. Specifically, AnEn correlation is better at about 300 sites out of 458, with cases where AnEn improves more than 100%. However, as shown in Fig. 6, overall the correlation for $PM_{2.5}$ is improved significantly with AnEn, which can be explained by the fact that often the degradation happens where $PM_{2.5}$ concentration are low (the latter happens at several locations, not shown).

Furthermore, we examine the performance of AnEn for extreme events by computing the correlation coefficient, bias, and RMSE for both CMAQ and AnEn mean time series of ozone and $PM_{2.5}$ using only the observations above the 95% quantile computed independently at each lead time over the verification period. The estimated bias and RMSE for extreme events are shown in Fig. 9. A lower RMSE and higher correlation coefficient of AnEn mean for both ozone and $PM_{2.5}$ at all the lead times shows that AnEn performs relatively well for the extreme events as well. However, the bias in AnEn is higher than CMAQ raw forecasts for extreme events of $PM_{2.5}$ mainly because of substantial reduction in the number of available quality analogs when we consider only extreme events. The RMSE can be decomposed in bias and centered root-mean-square error (CRMSE), the first being associated with systematic errors and the latter with conditional biases and random errors (Taylor 2001). A lower RMSE even at the lead times where AnEn bias is higher indicates that AnEn is reducing the random component of RMSE, CRMSE (i.e., exhibits a significantly improved predictive capability than CMAQ for rare events). Future work will focus on developing a bias correction technique to reduce the AnEn bias for the extreme events, similarly to what proposed by Alessandrini et al. (2019) for wind speed.

d. *Probabilistic predictions*

In this section, several attributes of AnEn and PeEn probabilistic predictions are evaluated to assess their performances for probabilistic AQ forecasting. These include the match between the observed and predicted cumulative PDF, reliability, resolution, statistical consistency, and an analysis of the spread–skill relationships (Jolliffe and Stephenson 2003; Wilks 2006).

1) OBSERVED VERSUS PREDICTED CUMULATIVE DISTRIBUTION

The Continuous ranked probability score is computed to assess the closeness between observed and predicted PDFs. The CPRS is calculated by comparing the full ensemble distribution with the observations, where both predictions and observations are represented as cumulative distribution functions (CDF; Carney and Cunningham 2006). It corresponds to the mean absolute error of deterministic predictions relative to the observations, and it has the same unit as the forecast variable. The more the ensemble-derived CDF is sharp and

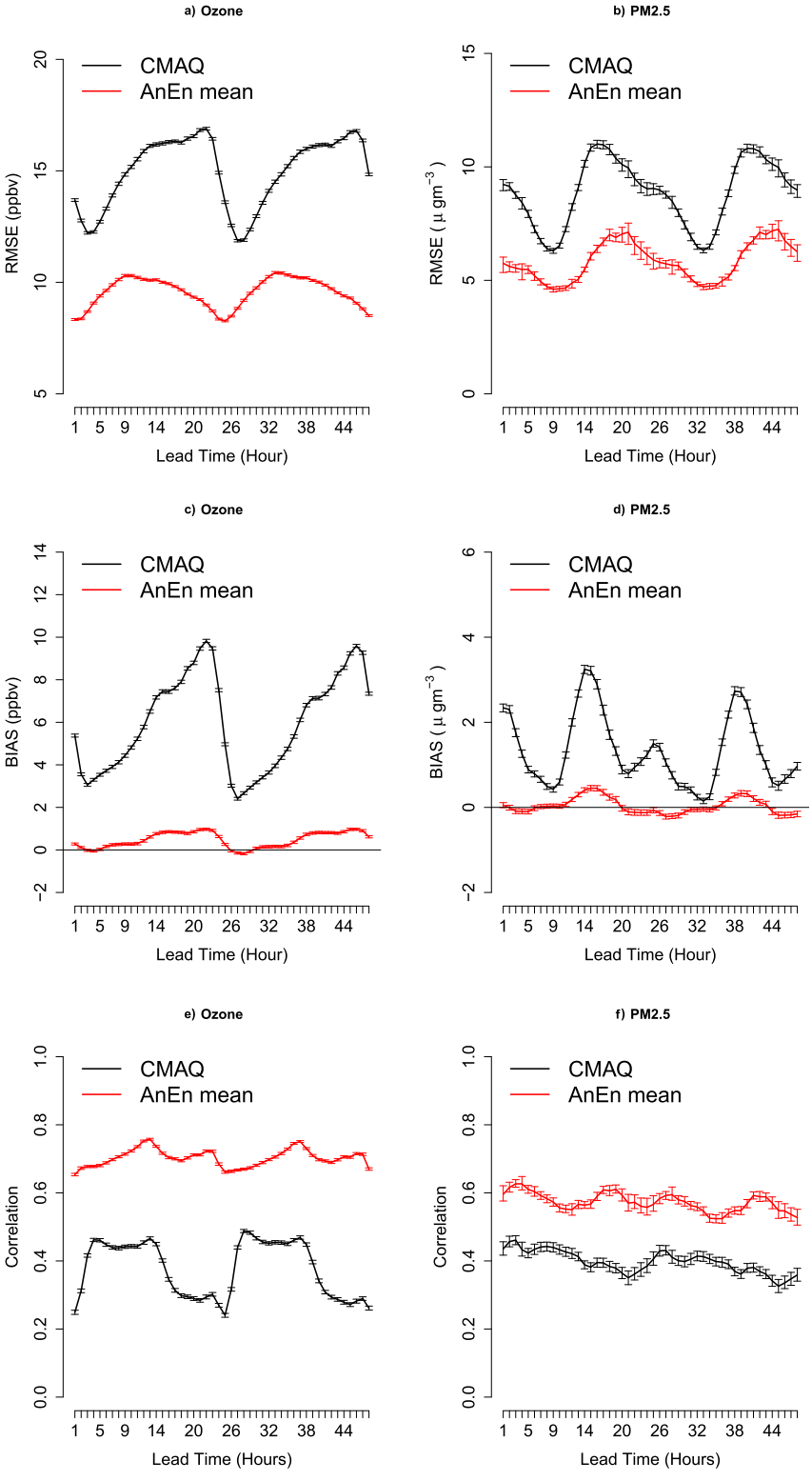

FIG. 6. (a),(b) RMSE, (c),(d) bias, and (e),(f) correlation for (left) O_3 and (right) $PM_{2.5}$ vs lead time in forecasts from CMAQ (black) and AnEn mean (red). Calculations are averages over all sites during the periods of study described in the text.

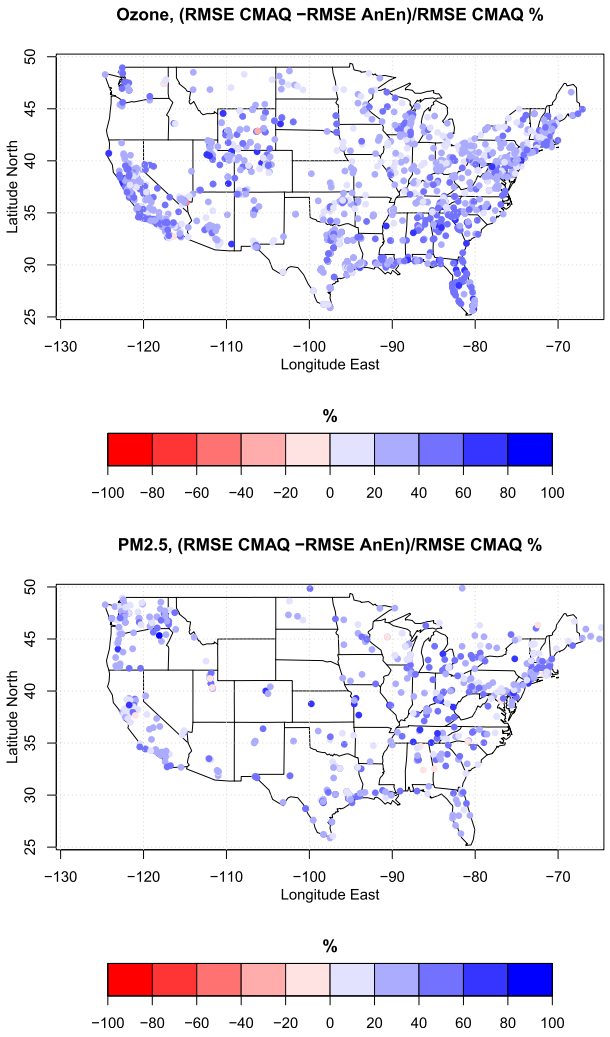

FIG. 7. Improvement (%; colors) to RMSE of forecasts of (top) O_3 and (bottom) $PM_{2.5}$ from AnEn vs CMAQ.

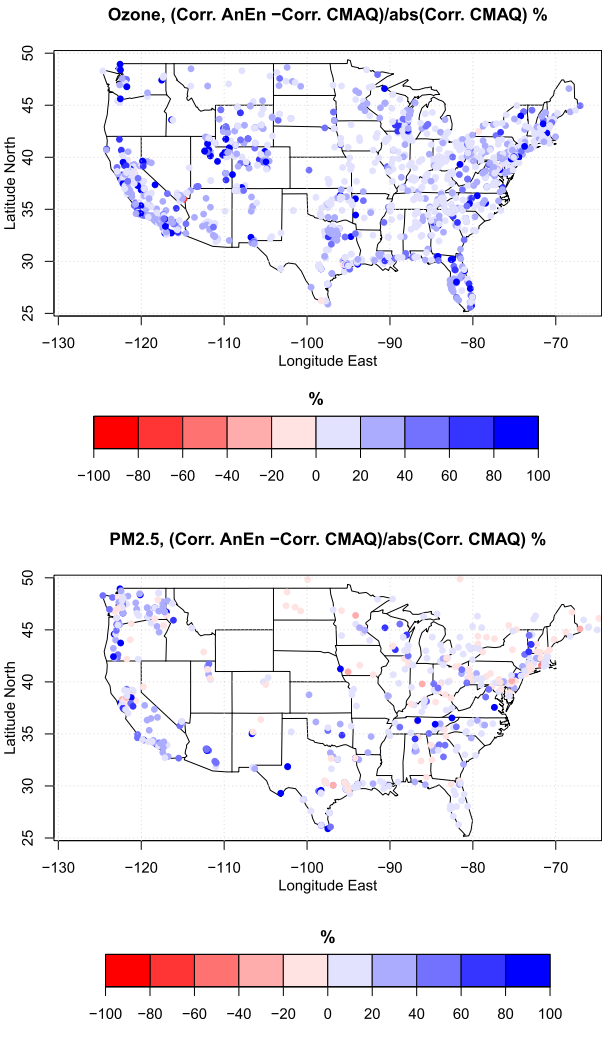

FIG. 8. As in Fig. 7, but for correlation.

centered on the corresponding observation, the lower the CRPS is. Zero is a perfect CRPS. CRPS can be decomposed in its components [i.e., CRPS = REL (reliability) + CRPSPOT, where CRPSPOT is the potential CRPS, with CRPSPOT = UNC − RES, where UNC is uncertainty and RES is resolution]. The potential CRPS is the forecasting system CRPS if it was perfectly reliable (REL = 0). For additional details on CRPS and its components formulation see Hersbach (2000).

As shown in Fig. 10, AnEn has a better (i.e., lower) CRPS than PeEn for most of the lead times of O_3 predictions, and all lead times of $PM_{2.5}$, indicating an overall better predictive probabilistic skill. The distinct diurnal cycle in each CRPS series is not surprising, given the diurnal cycle in forecast error (Fig. 6). The better CRPS of AnEn results from better resolution than PeEn (about 10% and 15% for O_3 and $PM_{2.5}$, respectively), an important attribute of probabilistic predictions. The reliability of the two systems is very similar. Reliability and resolution are discussed next.

2) RELIABILITY

An ensemble is reliable when its forecast probability matches the observed relative frequency (i.e., the rate of occurrence) of a certain condition over a long observational record. For instance, a reliable ensemble will predict a 7% probability of ground-level O_3 concentration exceeding a regulatory threshold at some point during a 24-h period if historically on 7% of days with similar conditions that threshold was exceeded. For a given condition (e.g., $O_3 > 50$ ppb), plotting forecast probabilities from a perfectly reliable model versus observed relative frequencies will result in a 1:1 diagonal line on a reliability diagram (Jolliffe and Stephenson 2003; Wilks 2006). However, the CRPS

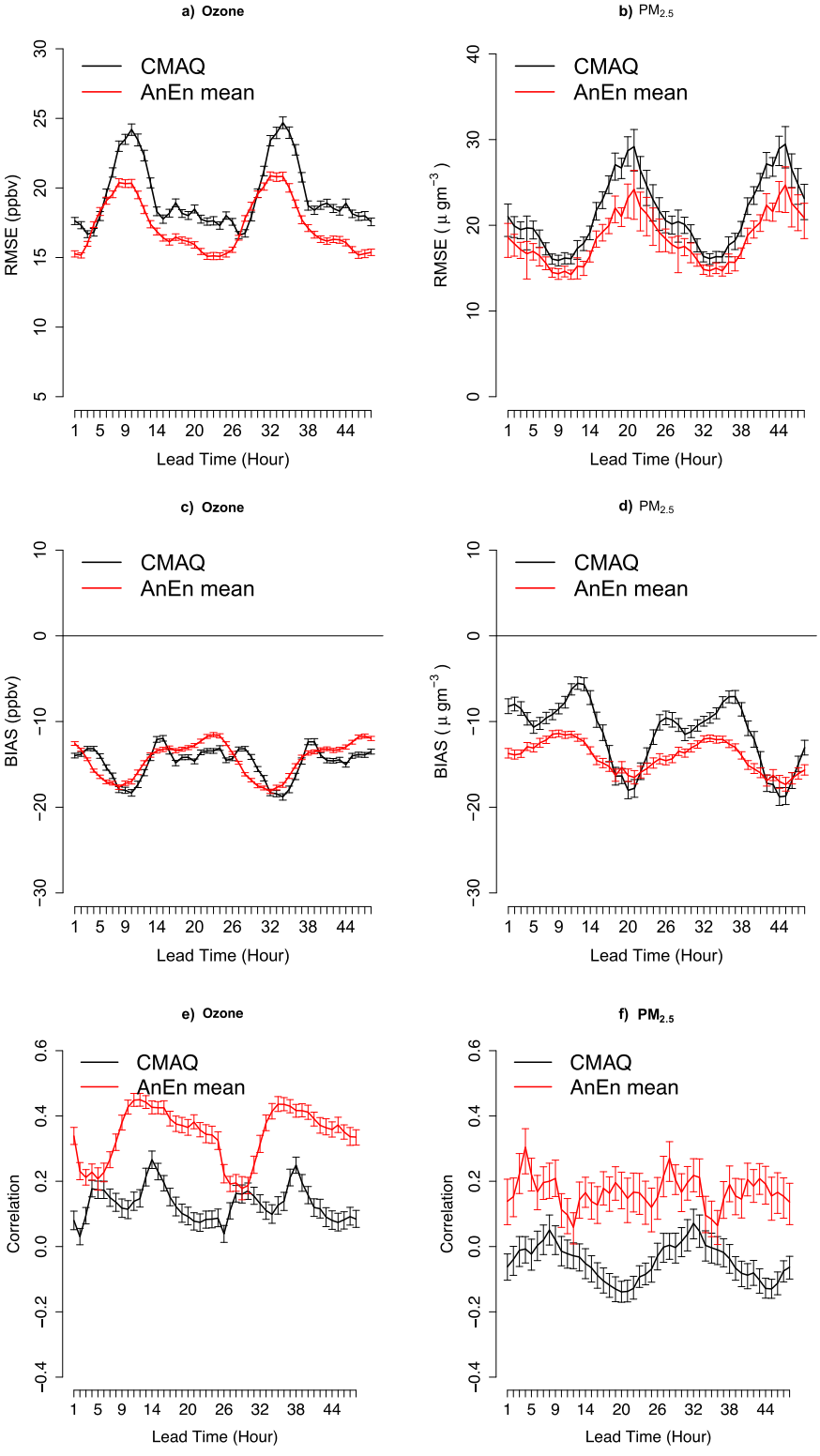

FIG. 9. As in Fig. 6, but with only the observations exceeding the 95% quantile computed independently over the verification period at each lead time and observation location.

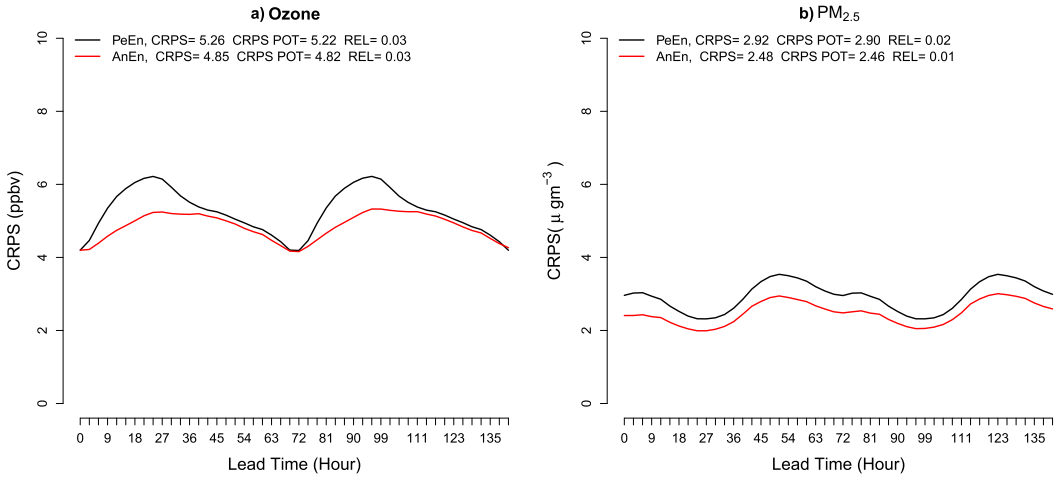

FIG. 10. CRPS vs lead time from forecasts of (a) O_3 and (b) $PM_{2.5}$ by AnEn (red) and PeEn (black).

reliability component provides an indication of overall reliability (i.e., not tied to a particular condition). For the latter, as shown in Fig. 10 and indicated by "REL," the reliability of AnEn and PeEn are very similar, and close to the perfect value of zero, indicating that both probabilistic prediction systems would provide an end-user with information that is not misleading (i.e., that would not lead to over or underconfidence in the forecast).

3) RESOLUTION

The potential CRPS ("POT" in Fig. 10) includes both uncertainty and resolution (Hersbach 2000). The uncertainty term is an attribute relative to the observations only and is the same across different prediction systems. However, the resolution quantifies the forecasts' ability to predict when an event occurs or not, or also the ability to separate different situations (i.e., the forecast system predicts a high/low values probability values correspondent to an occurring/not occurring event). Probability forecasts with perfect resolution are 100% on occasions when the event occurs and 0% when the event does not occur. The CRPS POT values reported in Fig. 10 show that AnEn has better resolution than PeEn. This is because AnEn forecasts are designed to capture errors in the current raw prediction, whereas PeEn includes the most recent 20 observations of O_3 or $PM_{2.5}$ at the same hour of day as the forecast valid time, which may not sample well the observation corresponding to the current prediction.

4) STATISTICAL CONSISTENCY

If an ensemble is statistically consistent, its members' forecasts are statistically indistinguishable from observations (Anderson 1996). If this condition is satisfied, when ranking an observation against the corresponding ensemble forecasts, the observation falls with equal probability in any of the ranks. Over a sufficient number of cases, when rank frequencies are plotted the resultant rank histogram is statistically flat if an ensemble is perfectly consistent (Anderson 1996; Hamill 2001; Talagrand et al. 1997). Rank histograms of forecasts from AnEn and PeEn are close to flat (Fig. 11). There is an overall slight high bias across all cases to both ensembles—shorter bars on the right of the histograms indicate that observations that fall among the higher predicted values are less common than those that fall among low values. The ensembles are also slightly underdispersive (except with AnEn for the ozone prediction) —the U shape at the tops of the bars indicate that observations that fall within the envelope of the ensembles' spreads are less common than those that fall outside the envelope, either lower than the lowest forecast value from a member or higher than the highest forecast value.

5) SPREAD–SKILL RELATIONSHIP

One way to assess an ensemble system ability to quantify the prediction uncertainty is by relating the spread (defined as the standard deviation of the members about the ensemble mean) among individual ensemble members' forecasts to the skill of their mean forecast, which is referred to as the spread–skill relationship (Delle Monache et al. 2013; van den Dool 1989). There are various ways to measure this relationship. Talagrand et al. (1997) reasoned that a statistically consistent ensemble's average standard deviation (a measure of spread) should match the RMSE of its mean forecast. Hopson (2014) provided insightful commentary on the topic.

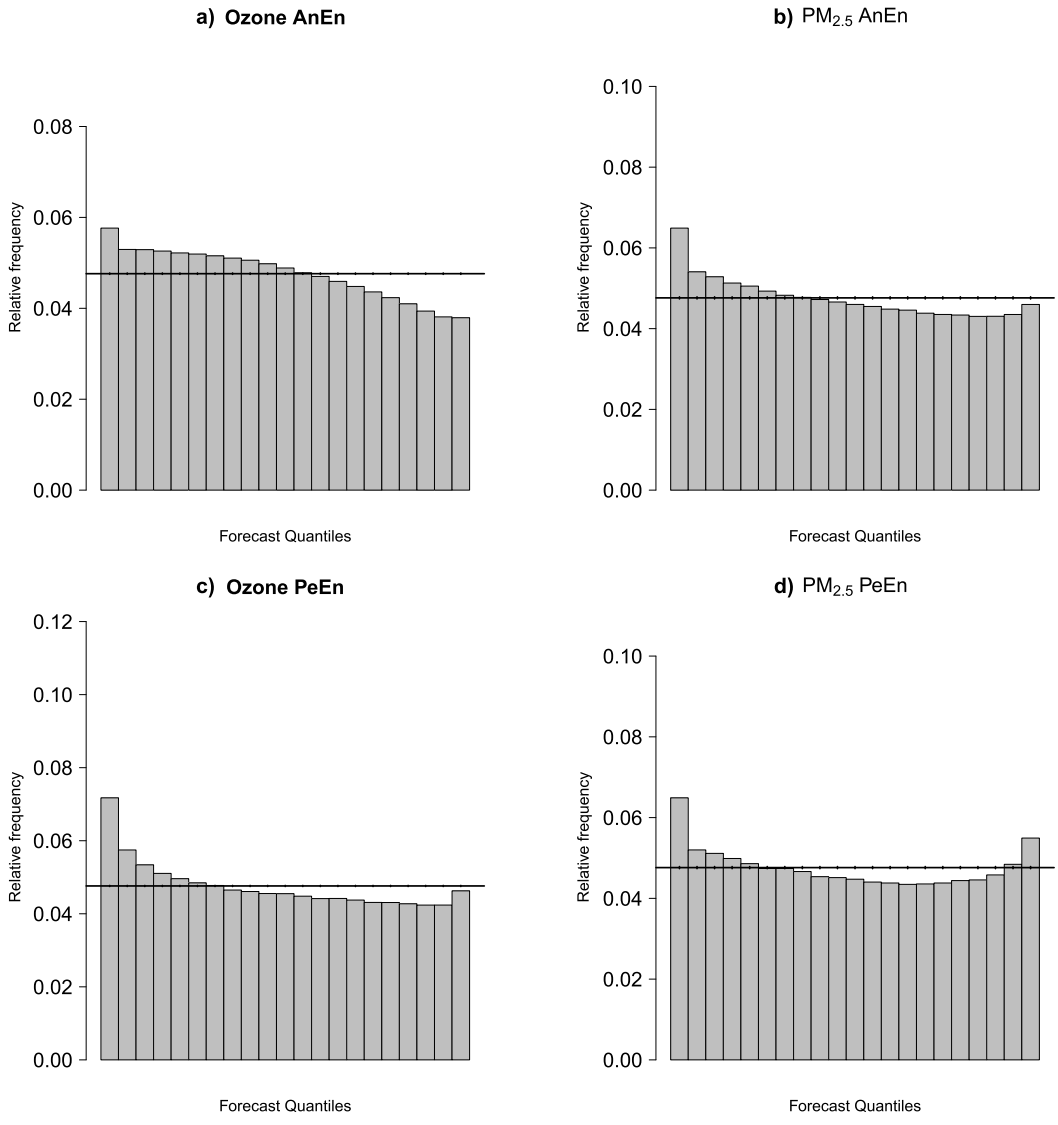

FIG. 11. Rank histograms of forecasts of (left) O_3 and (right) $PM_{2.5}$ from (a),(b) AnEn and (c),(d) the persistence ensemble. The horizontal line indicates the flat rank histogram that would result from a perfect ensemble, with vertical bars representing the confidence intervals for each bin.

We find that, indeed, standard deviation and RMSE correspond very well on average (across the different observational sites), especially for forecasts of $PM_{2.5}$ (Fig. 12). The correspondence is extremely robust over the 48-h forecast period, depending much more strongly on the diurnal cycle than on lead time. AnEn does not share PeEn's tendency to be underdispersive when forecasts of O_3 are less accurate, but forecasts of $PM_{2.5}$ from both models are similarly underdispersive.

One can also assess spread–skill relationship by examining model error versus spread after the latter is separated into bins that are subsets of the dataset's full range of spread. We find that forecasts from AnEn exhibit a strong spread–skill relationship according to this measure, as do forecasts from PeEn (Fig. 13). Both ensembles are slightly underdispersive when spread is small, which happens for the majority of bins, and slightly overdispersive when spread is large (evinced by the slopes $<$ 1:1 in Fig. 13), a feature that is more prominent for eastern stations than western ones, which is consistent with the U-shaped rank histograms in Fig. 11. The conditional bias displayed in Fig. 13 is relatively minor, however, and it might decrease with a larger archive of training cases (Delle Monache et al. 2013).

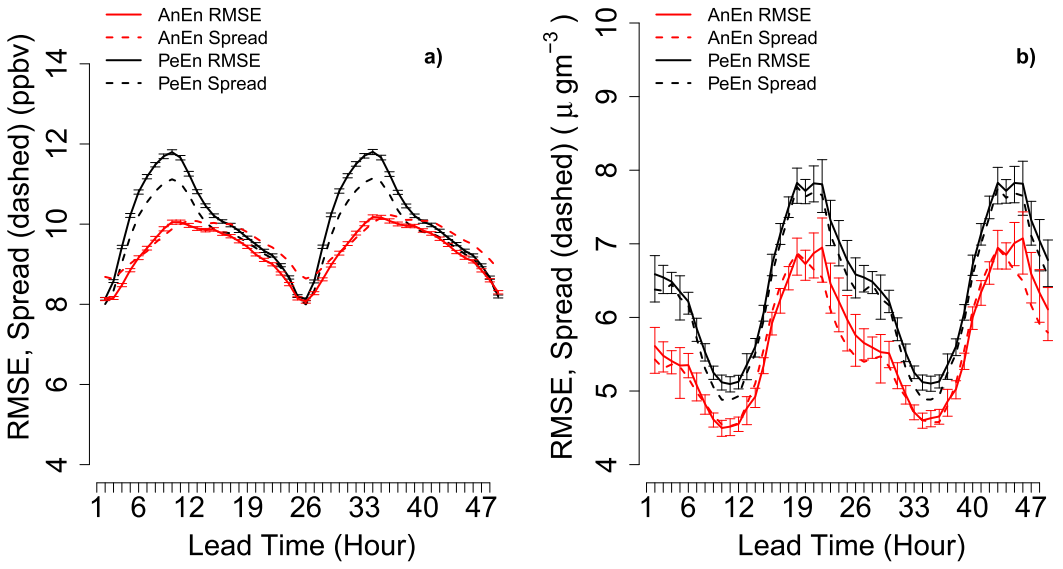

FIG. 12. Dispersion diagrams of forecasts of (a) O_3 (ppbv) and (b) $PM_{2.5}$ (μg m^{-3}) from the AnEn (red) and persistence ensemble (black). RMSE of the ensemble mean is solid; spread is dashed. Vertical bars span the 95% bootstrap confidence intervals.

4. Summary

Conventional air quality predictions are contaminated by uncertainties stemming from several sources, including initial conditions, emission, numerical approximations, and the simulation (or lack thereof) of physical and chemical processes. The ability to estimate these uncertainties in real time enhances decision-making to protect the public from poor air quality.

In this study, for the first time the analog ensemble technique has been applied to generate deterministic and probabilistic predictions of O_3 and $PM_{2.5}$. The available datasets span the period from 1 July 2014 to 30 September 2015 (456 days) for predictions of O_3 and from 1 July 2014 to 29 February 2016 (608 days) for predictions of $PM_{2.5}$. The verification periods to assess the performances of the predictive systems are June–September 2015 for O_3 and December 2015–February 2016 for $PM_{2.5}$. The analysis has been performed with 1045 and 458 stations for O_3 and $PM_{2.5}$, respectively, across the conterminous United States and southern Canada. The main findings are the following:

- The AnEn significantly improves the skill of deterministic predictions by reducing the errors of the deterministic model used to generate it, the Community Multiscale Air Quality, while increasing its correlation with the observations. For example, AnEn's root-mean-square error is lower than CMAQ's by roughly 35% and 30% for O_3 and $PM_{2.5}$, respectively, when computed across all sites and lead times.
- AnEn produces a probabilistic prediction that is statistically consistent, reliable, and sharp. It quantifies the uncertainty of the underlying prediction, which could contribute to an increased ability to protect the public health.
- An analog ensemble can be generated for existing real-time air quality forecast systems with very small additional computational cost in real time. However, an analog ensemble does require an archive of historical simulations from a deterministic model and observations of the quantity to be predicted, which can be built offline.

The results reported herein can be further improved with a longer training dataset (which would require additional computational resources), by extending existing training datasets to consider neighboring locations while searching analogs (at no additional computational cost), and by exploring more predictors for the analog-matching metric. Analog-based approaches, as other postprocessing methods, may exhibit limited skill when dealing with wildfire smoke events, particularly if those are not represented in the input file for emissions. That challenge can be addressed, at least partially, by including in the air quality prediction system a data assimilation component (e.g., assimilating aerosol optical depth from satellites, which has been shown to significantly improve the characterization of smoke from wildfires in air quality predictions) (Kumar et al. 2019).

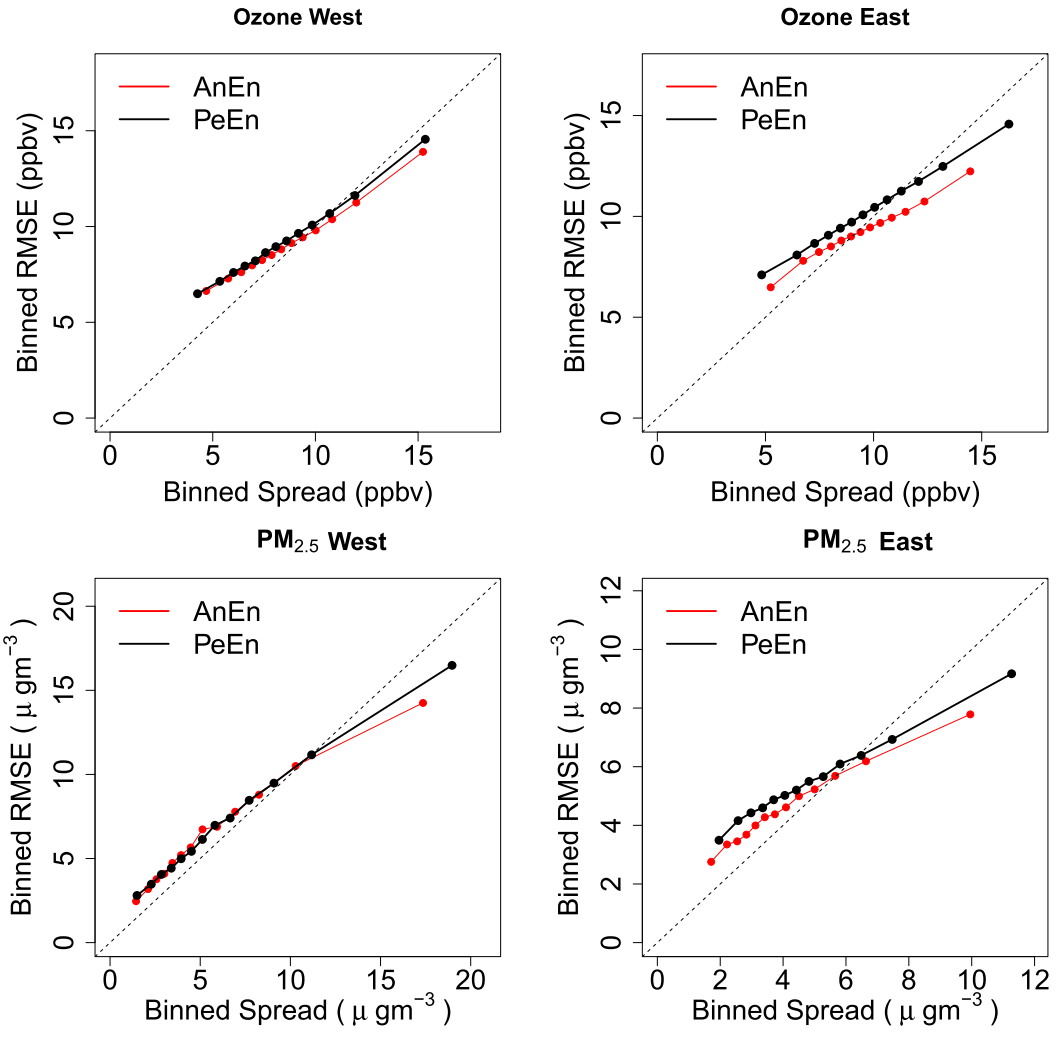

FIG. 13. RMSE vs spread of forecasts of (top) O_3 (ppbv) and (bottom) $PM_{2.5}$ ($\mu g\,m^{-3}$) for stations (left) west and (right) east of 100°W longitude for the AnEn (red) and persistence ensemble (black) calculated over bins with the same number of samples (dots).

Acknowledgments. This work was supported by NASA Earth Science Division Applied Science Program (Grant NNX15AH03G). We thank Jianping Huang (NCEP) for providing the CMAQ and Airnow datasets used in this analysis. The National Center for Atmospheric Research is sponsored by the National Science Foundation (NSF). The analog ensemble's code can be shared for research purposes upon request to the corresponding author. The datasets used in this study can be made available upon request to the corresponding author.

REFERENCES

Alessandrini, S., L. Delle Monache, S. Sperati, and G. Cervone, 2015a: An analog ensemble for short-term probabilistic solar power forecast. *Appl. Energy*, **157**, 95–110, https://doi.org/10.1016/j.apenergy.2015.08.011.

——, ——, ——, and J. N. Nissen, 2015b: A novel application of an analog ensemble for short-term wind power forecasting. *Renewable Energy*, **76**, 768–781, https://doi.org/10.1016/j.renene.2014.11.061.

——, ——, C. M. Rozoff, and W. E. Lewis, 2018: Probabilistic prediction of tropical cyclone intensity with an analog ensemble. *Mon. Wea. Rev.*, **146**, 1723–1744, https://doi.org/10.1175/MWR-D-17-0314.1.

——, S. Sperati, and L. Delle Monache, 2019: Improving the analog ensemble wind speed forecasts for rare events. *Mon. Wea. Rev.*, **147**, 2677–2692, https://doi.org/10.1175/MWR-D-19-0006.1.

Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate*, **9**, 1518–1530, https://doi.org/10.1175/1520-0442(1996)009<1518:AMFPAE>2.0.CO;2.

Barker, T. W., 1991: The relationship between spread and forecast error in extended-range forecasts. *J. Climate*, **4**, 733–742, https://doi.org/10.1175/1520-0442(1991)004<0733:TRBSAF>2.0.CO;2.

Bei, N., W. Lei, M. Zavala, and L. T. Molina, 2010: Ozone predictabilities due to meteorological uncertainties in the Mexico City basin using ensemble forecasts. *Atmos. Chem. Phys.*, **10**, 6295–6309, https://doi.org/10.5194/acp-10-6295-2010.

Buizza, R., 2008: The value of probabilistic prediction. *Atmos. Sci. Lett.*, **9**, 36–42, https://doi.org/10.1002/asl.170.

——, P. L. Houtekamer, G. Pellerin, Z. Toth, Y. Zhu, and M. Wei, 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Wea. Rev.*, **133**, 1076–1097, https://doi.org/10.1175/MWR2905.1.

Byun, D., and K. L. Schere, 2006: Review of the governing equations, computational algorithms, and other components of the Models-3 Community Multiscale Air Quality (CMAQ) modeling system. *Appl. Mech. Rev.*, **59**, 51, https://doi.org/10.1115/1.2128636.

Carmichael, G. R., A. Sandu, T. Chai, D. N. Daescu, E. M. Constantinescu, and Y. Tang, 2008: Predicting air quality: Improvements through advanced methods to integrate models and measurements. *J. Comput. Phys.*, **227**, 3540–3571, https://doi.org/10.1016/j.jcp.2007.02.024.

Carney, M., and P. Cunningham, 2006: Evaluating density forecasting models. Trinity College Dublin Dept. of Computer Science Tech. Rep. TCD-CS-2006-21, 12 pp.

Cervone, G., L. Clemente-Harding, S. Alessandrini, and L. Delle Monache, 2017: Short-term photovoltaic power forecasting using artificial neural networks and an analog ensemble. *Renewable Energy*, **108**, 274–286, https://doi.org/10.1016/j.renene.2017.02.052.

Dabberdt, W. F., and Coauthors, 2004: Meteorological research needs for improved air quality forecasting: Report of the 11th prospectus development team of the U.S. Weather Research Program. *Bull. Amer. Meteor. Soc.*, **85**, 563–586, https://doi.org/10.1175/BAMS-85-4-563.

Dalcher, A., E. Kalnay, and R. N. Hoffman, 1988: Medium range lagged average forecasts. *Mon. Wea. Rev.*, **116**, 402–416, https://doi.org/10.1175/1520-0493(1988)116<0402:MRLAF>2.0.CO;2.

Davò, F., S. Alessandrini, S. Sperati, L. Delle Monache, D. Airoldi, and M. T. Vespucci, 2016: Post-processing techniques and principal component analysis for regional wind power and solar irradiance forecasting. *Sol. Energy*, **134**, 327–338, https://doi.org/10.1016/j.solener.2016.04.049.

Delle Monache, L., 2010: Ensemble-based air quality predictions. *Air Quality: Theories, Methodologies, Computational Techniques, and Available Databases and Software*, Vol. IV, chapter 16C, (Advances and Updates), P. Zannetti, Ed., The EnviroComp Institute and the Air & Waste Management Association, 319–341.

——, and R. B. Stull, 2003: An ensemble air-quality forecast over western Europe during an ozone episode. *Atmos. Environ.*, **37**, 3469–3474, https://doi.org/10.1016/S1352-2310(03)00475-8.

——, X. Deng, Y. Zhou, and R. Stull, 2006a: Ozone ensemble forecasts: 1. A new ensemble design. *J. Geophys. Res.*, **111**, D05307, https://doi.org/10.1029/2005JD006310.

——, T. Nipen, X. Deng, Y. Zhou, and R. Stull, 2006b: Ozone ensemble forecasts: 2. A Kalman filter predictor bias correction. *J. Geophys. Res.*, **111**, D05308, https://doi.org/10.1029/2005JD006311.

——, J. P. Hacker, Y. Zhou, X. Deng, and R. B. Stull, 2006c: Probabilistic aspects of meteorological and ozone regional ensemble forecasts. *J. Geophys. Res.*, **111**, D24307, https://doi.org/10.1029/2005JD006917.

——, and Coauthors, 2008: A Kalman-filter bias correction method applied to deterministic, ensemble averaged and probabilistic forecasts of surface ozone. *Tellus*, **60B**, 238–249, https://doi.org/10.1111/j.1600-0889.2007.00332.x.

——, T. Nipen, Y. Liu, G. Roux, and R. Stull, 2011: Kalman filter and analog schemes to postprocess numerical weather predictions. *Mon. Wea. Rev.*, **139**, 3554–3570, https://doi.org/10.1175/2011MWR3653.1.

——, F. A. Eckel, D. L. Rife, B. Nagarajan, and K. Searight, 2013: Probabilistic weather prediction with an analog ensemble. *Mon. Wea. Rev.*, **141**, 3498–3516, https://doi.org/10.1175/MWR-D-12-00281.1.

Djalalova, I., and Coauthors, 2010: Ensemble and bias-correction techniques for air quality model forecasts of surface O3 and PM2.5 during the TEXAQS-II experiment of 2006. *Atmos. Environ.*, **44**, 455–467, https://doi.org/10.1016/j.atmosenv.2009.11.007.

——, L. Delle Monache, and J. Wilczak, 2015: PM2.5 analog forecast and Kalman filter post-processing for the Community Multiscale Air Quality (CMAQ) model. *Atmos. Environ.*, **108**, 76–87, https://doi.org/10.1016/j.atmosenv.2015.02.021.

Du, J., S. L. Mullen, and F. Sanders, 1997: Short-range ensemble forecasting of quantitative precipitation. *Mon. Wea. Rev.*, **125**, 2427–2459, https://doi.org/10.1175/1520-0493(1997)125<2427:SREFOQ>2.0.CO;2.

Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461–2480, https://doi.org/10.1175/1520-0493(2001)129<2461:AOAPMS>2.0.CO;2.

Ebisuzaki, W., and E. Kalnay, 1991: Ensemble experiments with a new lagged average forecasting scheme. WMO Research Activities in Atmospheric and Oceanic Modelling Rep. 15, 308 pp.

EPA, 2017: AIRNow-air quality monitor maps. Accessed 24 May 2017, https://www.airnow.gov/index.cfm?action=airnow.pointmaps.

Forouzanfar, M. H., and Coauthors, 2015: Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks in 188 countries, 1990–2013: A systematic analysis for the Global Burden of Disease Study 2013. *Lancet*, **386**, 2287–2323, https://doi.org/10.1016/S0140-6736(15)00128-2.

Galmarini, S., R. Bianconi, R. Bellasio, and G. Graziani, 2001: Forecasting the consequences of accidental releases of radionuclides in the atmosphere from ensemble dispersion modelling. *J. Environ. Radioact.*, **57**, 203–219, https://doi.org/10.1016/S0265-931X(01)00017-0.

——, and Coauthors, 2004: Ensemble dispersion forecasting—Part I: Concept, approach and indicators. *Atmos. Environ.*, **38**, 4607–4617, https://doi.org/10.1016/j.atmosenv.2004.05.030.

Garaud, D., and V. Mallet, 2010: Automatic generation of large ensembles for air quality forecasting using the Polyphemus system. *Geosci. Model Dev.*, **3**, 69–85, https://doi.org/10.5194/gmd-3-69-2010.

Hacker, J. P., and Coauthors, 2011: The U.S. Air Force Weather Agency's mesoscale ensemble: Scientific description and performance results. *Tellus*, **63A**, 625–641, https://doi.org/10.1111/j.1600-0870.2010.00497.x.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550–560, https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2.

——, and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Wea. Rev.*, **134**, 3209–3229, https://doi.org/10.1175/MWR3237.1.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*,

15, 559–570, https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Hoffman, R. N., and E. Kalnay, 1983: Lagged average forecasting, an alternative to Monte Carlo forecasting. *Tellus*, **35A**, 100–118, https://doi.org/10.3402/TELLUSA.V35I2.11425.

Hopson, T. M., 2014: Assessing the ensemble spread–error relationship. *Mon. Wea. Rev.*, **142**, 1125–1142, https://doi.org/10.1175/MWR-D-12-00111.1.

Huang, J., and Coauthors, 2017: Improving NOAA NAQFC $PM_{2.5}$ predictions with a bias correction approach. *Wea. Forecasting*, **32**, 407–421, https://doi.org/10.1175/WAF-D-16-0118.1.

Jolliffe, I. T., and D. B. Stephenson, Eds., 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley and Sons, 254 pp.

Junk, C., L. Delle Monache, S. Alessandrini, G. Cervone, and L. von Bremen, 2015: Predictor-weighting strategies for probabilistic wind power forecasting with an analog ensemble. *Meteor. Z.*, **24**, 361–379, https://doi.org/10.1127/metz/2015/0659.

Kalnay, E., 2003: *Atmospheric Modeling, Data Assimilation and Predictability*. 1st ed. Cambridge University Press, 341 pp.

Kioutsioukis, I., and Coauthors, 2016: Insights into the deterministic skill of air quality ensembles from the analysis of AQMEII data. *Atmos. Chem. Phys.*, **16**, 15 629–15 652, https://doi.org/10.5194/acp-16-15629-2016.

Krishnamurti, T. N., 1999: Improved weather and seasonal climate forecasts from multimodel superensemble. *Science*, **285**, 1548–1550, https://doi.org/10.1126/science.285.5433.1548.

Kumar, R., and Coauthors, 2019: Toward improving short-term predictions of fine particulate matter over the United States via assimilation of satellite aerosol optical depth retrievals. *J. Geophys. Res. Atmos.*, **124**, 2753–2773, https://doi.org/10.1029/2018JD029009.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, **102**, 409–418, https://doi.org/10.1175/1520-0493(1974)102<0409:TSOMCF>2.0.CO;2.

Lu, C., H. Yuan, B. E. Schwartz, and S. G. Benjamin, 2007: Short-range numerical weather prediction using time-lagged ensembles. *Wea. Forecasting*, **22**, 580–595, https://doi.org/10.1175/WAF999.1.

Ma, J., Y. Zhu, R. Wobus, and P. Wang, 2012: An effective configuration of ensemble size and horizontal resolution for the NCEP GEFS. *Adv. Atmos. Sci.*, **29**, 782–794, https://doi.org/10.1007/s00376-012-1249-y.

Mallet, V., 2010: Ensemble forecast of analyses: Coupling data assimilation and sequential aggregation. *J. Geophys. Res.*, **115**, D24303, https://doi.org/10.1029/2010JD014259.

——, and B. Sportisse, 2006a: Ensemble-based air quality forecasts: A multimodel approach applied to ozone. *J. Geophys. Res.*, **111**, D18302, https://doi.org/10.1029/2005JD006675.

——, and ——, 2006b: Uncertainty in a chemistry-transport model due to physical parameterizations and numerical approximations: An ensemble approach applied to ozone modeling. *J. Geophys. Res.*, **111**, D01302, https://doi.org/10.1029/2005JD006149.

——, A. Nakonechny, and S. Zhuk, 2013: Minimax filtering for sequential aggregation: Application to ensemble forecast of ozone analyses. *J. Geophys. Res. Atmos.*, **118**, 11 294–11 303, https://doi.org/10.1002/JGRD.50751.

Marécal, V., and Coauthors, 2015: A regional air quality forecasting system over Europe: The MACC-II daily ensemble production. *Geosci. Model Dev.*, **8**, 2777–2813, https://doi.org/10.5194/gmd-8-2777-2015.

McKeen, S., and Coauthors, 2005: Assessment of an ensemble of seven real-time ozone forecasts over eastern North America during the summer of 2004. *J. Geophys. Res.*, **110**, D21307, https://doi.org/10.1029/2005JD005858.

——, and Coauthors, 2007: Evaluation of several $PM_{2.5}$ forecast models using data collected during the ICARTT/NEAQS 2004 field study. *J. Geophys. Res.*, **112**, D10S20, https://doi.org/10.1029/2006JD007608.

Mittermaier, M. P., 2007: Improving short-range high-resolution model precipitation forecast skill using time-lagged ensembles. *Quart. J. Roy. Meteor. Soc.*, **133**, 1487–1500, https://doi.org/10.1002/qj.135.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73–119, https://doi.org/10.1002/qj.49712252905.

Monteiro, A., and Coauthors, 2013: Bias correction techniques to improve air quality ensemble predictions: Focus on O_3 and PM over Portugal. *Environ. Model. Assess.*, **18**, 533–546, https://doi.org/10.1007/s10666-013-9358-2.

Muller, N. Z., and R. Mendelsohn, 2007: Measuring the damages of air pollution in the United States. *J. Environ. Econ. Manage.*, **54**, 1–14, https://doi.org/10.1016/j.jeem.2006.12.002.

Murphy, J. M., 1988: The impact of ensemble forecasts on predictability. *Quart. J. Roy. Meteor. Soc.*, **114**, 463–493, https://doi.org/10.1002/qj.49711448010.

Nagarajan, B., L. Delle Monache, J. P. Hacker, D. L. Rife, K. Searight, J. C. Knievel, and T. N. Nipen, 2015: An evaluation of analog-based post-processing methods across several variables and forecast models. *Wea. Forecasting*, **30**, 1623–1643, https://doi.org/10.1175/WAF-D-14-00081.1.

Pagowski, M., and Coauthors, 2005: A simple method to improve ensemble-based ozone forecasts. *Geophys. Res. Lett.*, **32**, L07814, https://doi.org/10.1029/2004GL022305.

Palmer, T. N., 2002: The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Quart. J. Roy. Meteor. Soc.*, **128**, 747–774, https://doi.org/10.1256/0035900021643593.

Potempski, S., and S. Galmarini, 2009: Est modus in rebus: Analytical properties of multi-model ensembles. *Atmos. Chem. Phys.*, **9**, 9471–9489, https://doi.org/10.5194/acp-9-9471-2009.

——, and Coauthors, 2008: Multi-model ensemble analysis of the ETEX-2 experiment. *Atmos. Environ.*, **42**, 7250–7265, https://doi.org/10.1016/j.atmosenv.2008.07.027.

Solazzo, E., and Coauthors, 2012: Model evaluation and ensemble modelling of surface-level ozone in Europe and North America in the context of AQMEII. *Atmos. Environ.*, **53**, 60–74, https://doi.org/10.1016/j.atmosenv.2012.01.003.

Sperati, S., S. Alessandrini, and L. Delle Monache, 2017: Gridded probabilistic weather forecasts with an analog ensemble. *Quart. J. Roy. Meteor. Soc.*, **143**, 2874–2885, https://doi.org/10.1002/QJ.3137.

Talagrand, O., R. Vautard, and B. Strauss, 1997: Evaluation of probabilistic prediction systems. *Proc. ECMWF Workshop on Predictability*, Reading, United Kingdom, ECMWF, 1–26.

Taylor, K. E., 2001: Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.*, **106**, 7183–7192, https://doi.org/10.1029/2000JD900719.

Toth, Z., and E. Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, **125**, 3297–3319, https://doi.org/10.1175/1520-0493(1997)125<3297:EFANAT>2.0.CO;2.

van den Dool, H. M., 1989: A new look at weather forecasting through analogues. *Mon. Wea. Rev.*, **117**, 2230–2247, https://doi.org/10.1175/1520-0493(1989)117<2230:ANLAWF>2.0.CO;2.

Vautard, R., and Coauthors, 2012: Evaluation of the meteorological forcing used for the Air Quality Model Evaluation International Initiative (AQMEII) air quality simulations. *Atmos. Environ.*, **53**, 15–37, https://doi.org/10.1016/j.atmosenv.2011.10.065.

Wilks, D. S., 2006: *Statistical Methods in the Atmopsheric Sciences*. 2nd ed. International Geophysics Series, Vol. 100, Academic Press, 648 pp.

Žabkar, R., D. Koračin, and J. Rakovec, 2013: A WRF/Chem sensitivity study using ensemble modelling for a high ozone episode in Slovenia and the Northern Adriatic area. *Atmos. Environ.*, **77**, 990–1004, https://doi.org/10.1016/j.atmosenv.2013.05.065.

Zhang, F., N. Bei, J. W. Nielsen-Gammon, G. Li, R. Zhang, A. Stuart, and A. Aksoy, 2007: Impacts of meteorological uncertainties on ozone pollution predictability estimated through meteorological and photochemical ensemble forecasts. *J. Geophys. Res.*, **112**, D04304, https://doi.org/10.1029/2006JD007429.

Zhang, Y., M. Bocquet, V. Mallet, C. Seigneur, and A. Baklanov, 2012: Real-time air quality forecasting, part I: History, techniques, and current status. *Atmos. Environ.*, **60**, 632–655, https://doi.org/10.1016/j.atmosenv.2012.06.031.