Enhancing air quality forecasts across the contiguous United States (CONUS) during wildfires using Analog-based post-processing methods

Maryam Golbazi[1], Stefano Alessandrini[1], Rajesh Kumar[1], Paddy McCarthy[1], Patrick C. Campbell[2,4], Piyush Bhardwaj[1], Cenlin He[1], and Jeffery McQueen[3]

[1]Research Application Laboratory (RAL), National Center for Atmospheric Research (NCAR), Boulder, CO 80301, USA.
[2]Air Resources Laboratory, NOAA, NCWCP, 5830 University Research Ct., College Park, MD 20740, USA.
[3]Environmental Modeling Center (EMC), NCEP, NWS, NOAA, 5830 University Research Ct., College Park, MD 20740, USA.
[4]Cooperative Institute for Satellite Earth System Studies (CISESS), Center for Spatial Information Science and System (CSISS), George Mason University, 4400 University Drive, Fairfax, VA 22030, USA.

Correspondence: Maryam Golbazi, Research Application Laboratory, National Center for Atmospheric Research, Boulder, CO 80301, USA (mgolbazi@ucar.edu)

Abstract.

With our growing understanding of the risks of air pollution to human health, air quality forecasting has become a very important tool to enable decision-makers to take preventive and corrective measures for current and future policies. In addition, accurate predictions of air quality can help predict and mitigate the impacts of wildfires on human health, which have an increased risk due to anthropogenic climate change. However, errors in air quality forecasts limit their value in decision-making processes. Thus, increasing the accuracy of air quality forecasts is of significant importance. In this study, we have utilized the Community Multiscale Air Quality (CMAQ) modeling system with a 12 km horizontal grid resolution to generate daily 48-hr fine particulate matter ($PM_{2.5}$) forecasts for the Contiguous United States (CONUS) domain for June 1st through September 29th (major wildfire season) during 2015-2021. We conduct CMAQ offline simulations using meteorological inputs generated by the NOAA's Unified Forecast System (UFS) numerical weather prediction model. We have also included a Carbon Monoxide-FIRE ($table$) tracer in CMAQ, which tracks CO emitted by wildfires. we analyze the performance of the CMAQ $PM_{2.5}$ and UFS meteorological forecasts over seven years of simulations for Environmental Protection Agency (EPA)-defined ten regions using the Air Quality System (AQS) ambient air pollution data from over a thousand monitoring sites across the CONUS. We have found that on average, the CMAQ model performs better in the eastern CONUS with the lowest Root Mean Squared Error (RMSE) (7-14 $\mu g/m^3$) while in the west, where wildfires are prevalent, the model has the highest RMSE of up to 35 $\mu g/m^3$. Next, we employ the state-of-the-art Analog Ensemble (AnEn) method to improve the accuracy of the forecasts and quantify the forecast improvements by AnEn. We also introduce two new predictors, i.e., a $CO-FIRE$ tracer in the model, and the maximum observed $PM_{2.5}$ concentrations from the previous day ($PM_{2.5}-preday-max$). Despite the challenges of using AnEn for wildfires, we demonstrate that it has the potential to improve the CMAQ model forecast over the CONUS. We find that AnEn decreases the model RMSE by up to 25%, including additional 7% and 15% reduction by $CO-FIRE$ and $PM_{2.5}-preday-max$ predictors, respectively, at different forecast lead times. In addition, the correlation between AnEn

forecasts and observations is 20%-40% higher than that between CMAQ and observations. The Mean Bias Error (MBE) for AnEn forecasts is consistent and approximately -0.5 $\mu g/m^3$ whereas CMAQ MBE varies between -1 and +1 $\mu g/m^3$ between 0-48 forecast hours. AnEn significantly improves the PM2.5 forecast results during its highest episodes. During the initial phases of wildfires, AnEn performs similarly to CMAQ. However, it soon catches up and decreases the error significantly.

keywords: air quality forecasting, wildfires, analog ensemble, bias correction,

1 Abbreviations

AnEn = Analog Ensemble.

AQS = Air Quality System.

BEIS = Biogenic Emission Inventory System.

BCON = Boundary Conditions Processor. CO = Carbon Monoxide.

CMAQ = Community Multiscale Air Quality model.

CSI = Critical Success Index.

CONUS = Continuous United States.

EPA = Environmental Protection Agency.

FINN = Fire Inventory from NCAR.

FV3 = Finite Volume Cubed

ICON = Initial Conditions Processor.

MAE = Mean Absolute Error.

MBE = Mean Bias Error.

MRW App = Medium-Range Weather Application.

NACC = NOAA-EPA Atmosphere-Chemistry Coupler.

NCAR = National Center for Atmospheric Research.

NEI = National Emission Inventory.

NOAA = National Oceanic and Atmospheric Administration.

$PM_{2.5}$ = Fine Particulate Matter with a size smaller than 2.5 $\mu g/m^3$.

$PM_{2.5} - model$ = $PM_{2.5}$ forecast by the model.

$PM_{2.5} - preday - max$ = maximum observed $PM_{2.5}$ from the previous day measurements.

RMSE = Root Mean Squared Error.

SRW App = Short-Range Weather Application.

UFS = Unified Forecast System.

WACCM = Whole Atmosphere Community Climate Model.

2 Introduction

Air quality predictions provide decision-makers with a valuable tool to mitigate various risks associated with poor air quality. Nonetheless, the usefulness of these forecasts in the decision-making process may be limited due to the uncertainties involved in the predictions. Complete elimination of uncertainty in air-quality forecasting is not feasible. Nevertheless, there are effective methods to address the inevitable uncertainty and minimize its impact. Uncertainties in air quality forecasting can originate from meteorological inputs (Kumar et al. (2019); Ryu et al. (2018); Zhang et al. (2007)), numerical noise in the model (Ancell et al. (2018); Golbazi et al. (2022)), numerical approximations, errors in emission inputs (Foley et al. (2015)), etc.

One way to characterize the uncertainty is to use ensembles instead of a single forecast model. Ensembles offer several advantages; they provide probabilistic guidance that can be significantly more valuable for decision-making than a single forecast (Buizza (2008); Palmer (2002)). An ensemble's mean forecast generally exhibits greater accuracy than any individual member's prediction (Du et al. (1997); Ebert (2001); Galmarini et al. (2001); Djalalova et al. (2010)). McKeen et al. (2007) reported that a simple average of six models for the $PM_{2.5}$ forecast referred to as the ensemble forecast outperformed each individual model. Similarly, in the 2006 Texas Air Quality field campaign, Djalalova et al. (2010) found that combining Kalman filtering with weighted model averaging led to a more accurate forecast, resulting in a 43% decrease in the RMSE and a 62% increase in the correlation coefficient when compared to using other methods to generate ensembles or individual models.

Probabilistic forecasts of weather variables can be generated using an ensemble of model runs with the members being different models (i.e., multi-model ensemble) (Zemouri et al. (2019); Ziehmann (2000)), having different initial conditions (Toth and Kalnay (1993); Molteni et al. (1996)), physics configurations (i.e., multi-physics) (Stensrud et al. (2000)), and stochastic perturbations of the tendencies of physics parameterizations (Buizza et al. (1999)). In these cases, the uncertainty of predicting a meteorological variable is represented by the ensemble spread, defined as the standard deviation of the members about the ensemble mean.

Alternatively, in a relatively new approach, the members of an ensemble are defined using statistical post-processing techniques such as the analog ensemble (AnEn) (Alessandrini et al. (2023, 2015); Hamill and Whitaker (2006); Delle Monache et al. (2013)). Delle Monache et al. (2011) developed a probabilistic weather prediction method using an Analog Ensemble (AnEn). The technique involves selecting historical weather patterns that are similar to the current conditions and using them to forecast future weather. The authors showed that this approach can capture the uncertainty of the forecasts and improve their accuracy, especially in regions with complex weather patterns while saving significant computational time.

Analog ensemble estimates the probability distribution of future atmospheric conditions by comparing past observations that best match the current model forecasts (Delle Monache et al. (2020)). For instance, in Delle Monache et al. (2013), AnEn is used to make probabilistic predictions of wind speed and temperature over the contiguous United States, compared to observations from surface stations, and evaluated against a state-of-the-science Numerical Weather Prediction ensemble system. The study finds that AnEn is consistent, reliable, captures flow-dependent behavior of errors, and performs similarly or better than other methods, such as logistic regression and ensemble model output statistics. Meanwhile, the AnEn has lower computational costs in real-time applications.

Recently, Alessandrini et al. (2018) successfully applied the AnEn method to tropical cyclone applications by constructing an AnEn from a database of Hurricane Weather Research and Forecast model forecasts and demonstrated that the forecasts of maximum sustained wind could be improved using a set of 6–8 predictor. Lewis et al. (2021) extended this work and derived an AnEn for the more specialized application of predicting tropical cyclone rapid intensity change, illustrating its effectiveness during a real-time test of the 2017 and 2018 Atlantic and Eastern Pacific hurricane seasons. In other cases where the unpredictability of meteorological variables, such as total cloud cover, makes accurate solar power predictions crucial (Alessandrini et al. (2023); Alessandrini (2022)), the use of an analog ensemble method generated probabilistic solar power forecasts based on historical data sets. The AnEn method determined performance as well as the quantile regression technique for common events while exhibiting better performance for rare events and during hours with a low solar elevation.

Meanwhile, in a review of the use of analogs to forecast the atmosphere, Weigel et al. (2008) discussed various methods for identifying historical weather patterns that are similar to the current conditions and using them to generate forecast ensembles. They showed that this approach can improve the accuracy of weather forecasts and provide valuable information on the uncertainty of the predictions.

As discussed above, several studies have demonstrated the AnEn's adaptability to a wide range of applications. These articles demonstrate the potential of analog ensembles in weather forecasting and offer valuable insights into their implementation and calibration. By using historical weather patterns to generate forecast ensembles, it is possible to improve the accuracy and reliability of weather forecasts and better estimate their uncertainty. Recently, we extended the application of the AnEn algorithm to 48-hr daily air quality forecasts from the Community Multi-scale Air Quality (CMAQ) model and found that AnEn can drastically improve the accuracy of air quality forecasts under normal conditions over the CONUS (Delle Monache et al. (2020)). Here, we explore if AnEn can improve the accuracy of fine particulate matter ($PM_{2.5}$) forecasts during wildfires, which is a challenging task owing to large interannual variability in fire emissions that makes it hard to find analogous fire-affected conditions in the past. $PM_{2.5}$ is a harmful air pollutant that consists of microscopic particles that can penetrate human lungs and even the bloodstream and cause serious health problems (U.S. Environmental Protection Agency (EPA) (2020)). It is one of the "criteria" pollutants that are regulated at the federal level by the U.S. EPA via the National Ambient Air Quality Standards (U.S. Environmental Protection Agency (EPA) (2022a)). We focus on the verification of the AnEn's performance over the EPA-defined ten regions to assess regional variability in AnEn performance as a function of varying wildfire influences. In order to monitor the dispersion of wildfire smoke throughout the domain, we have incorporated a $CO-FIRE$ tracer into the model. This tracer effectively traces the CO emitted by wildfires within the study area, serving as a reliable indicator of the smoke's spread from the wildfires across the CONUS.

3 Methods

Air quality forecasts used in this study are based on the UFS-CMAQ modeling system. CMAQ is a Cartesian air quality model that simulates the concentrations of atmospheric pollutants at regional scales using meteorological inputs. We perform our simulations and analysis in the domain of the contiguous United States (CONUS) (Fig. 1). Here, the meteorological inputs are

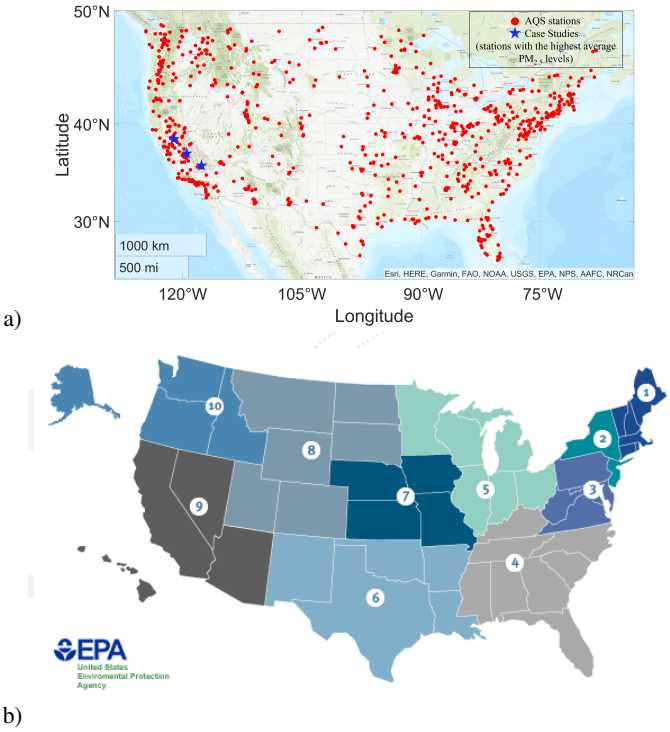

Figure 1. Domain of the study (CONUS). a) The red dots are the observing stations from which we have used the observational data. The three stations in blue are the stations with the highest average $PM_{2.5}$ levels over the study period (an indicator of wildfires), and b) the ten EPA-defined regions across the CONUS (https://www.nalms.org/2021nmc/12th-national-monitoring-conference-open-access-sessions/).

provided using the National Oceanic and Atmospheric Administration (NOAA) Unified Forecasting System (UFS) (Campbell et al. (2022)). $PM_{2.5}$ forecasts at 0–48 hour lead times are created for 7 major wildfire seasons (June-September) of 2015 - 2021.

We perform model performance analysis on the UFS-CMAQ modeling system. The model is evaluated against near-surface
125 atmospheric concentrations of $PM_{2.5}$ observed by the Air Quality System (AQS) network. Hourly averaged modeled $PM_{2.5}$ and meteorological parameters were paired in space and time with the hourly AQS observational data. Three statistical metrics are used to compare the observed and predicted hourly $PM_{2.5}$ concentrations. The Mean Bias Error (MBE) is used to measure model bias. The RMSE is used as a measure of model random errors. The correlation coefficient is used to represent the covariation between the model and observations.

Table 1. Details of the UFS-CMAQ model setup.

Simulation period	June 1st – Sept 30; 2015 – 2021
Horizontal grid resolution	12 km
Vertical layers	35
UFS MRW App v1.0	UFS SRW App v2.1
Initial/boundary conditions	Global Forecasting System (GFS) 6-hourly, 36-km resolution
LSM	Noah-modified 21-category IGBP-MODIS
Shortwave radiation	RRTMG shortwave
Longtwave radiation	RRTMG scheme
Grid size	442 × 265 grid cells
CMAQ version 5.3.2	
Chemistry	Carbon bond 6 revision 3
Aerosole module	AERO7
Meteorological inputs	UFS model
Anthropogenic emission data	EPA/NEI 2017
Biogenic emission data	BEIS
Fire emission data	FINN v2.2
Initial/boundary conditions	ICON/BCON
Grid size	315 × 300 grid cells

3.1 UFS-CMAQ modeling system

The meteorological fields that offline drive the CMAQ model are generated at a horizontal grid spacing of 12 x 12 km2 using the Unified Forecast System (UFS; https://ufscommunity.org/) Medium-Range (version 1.0) and Short-Range (version 2.1) Weather Applications. UFS is a community-based, coupled, comprehensive Earth modeling system developed by NOAA. UFS can be configured in multiple applications depending on the spatial and temporal scale of the problem. UFS Medium-Range Weather Application (MRW App) allows prediction of global weather behavior while the Short-Range Weather Application (SRW App) allows predictions of regional weather behavior. Since the SRW App was not available at the beginning of this project, we simulated June-September of 2017-2020 using the MRW App. However, the large computational costs associated with running a global MRW App prevented us from running it for other years. Since Analog Ensemble is expected to work better with longer-term training datasets, the release of the SRW App in the latter half of this project allowed us to simulate and include June-Sep of 2015, 2016, and 2021 in our study period. The UFS output cannot be directly supplied to CMAQ because UFS output is available on a Finite Volume Cubed Sphere (FV3) grid and CMAQ uses a Lambert conformal projection. Therefore, UFS output is mapped to the CMAQ domain through the NOAA-EPA Atmosphere-Chemistry Coupler (NACC) developed by Campbell et al. (2022). The CMAQ version 5.3.2 (U.S. Environmental Protection Agency (EPA) (2022b)) is configured with a horizontal

grid spacing of 12 x 12 km2 with 442, 265 horizontal grid points in the longitudinal, and latitudinal directions, respectively, and 35 levels in vertical. The model vertical grid stretches from the surface to 50 hPa. We employ the "$cb6r3 - ae7 - aq$" chemical mechanism that uses Carbon Bond 6 version r3 for gas-phase chemistry and the AERO7 aerosol module for representing aerosol processes including secondary organic aerosols. Anthropogenic emissions for the year 2017 are based on the National Emission Inventory (NEI) for 2017 and are derived for the other years by applying EPA-reported annual state-wise trends to the NEI 2017 emissions (https://www.epa.gov/air-emissions-inventories/2017-national-emissions-inventory-nei-data). This emission inventory contains gridded 2D emissions that are released into each grid cell of the modeling domain near the surface (i.e., "area sources", such as traffic or residential heating) and stack-specific "point sources", where each stack is assigned unique coordinates and parameters (i.e., smokestacks or ship chimneys), at a 12 km resolution. The anthropogenic emissions are distributed on a 442 times 265 grid, which covers the entire CONUS, with 35 layers in vertical. Fire emissions in CMAQ are represented using the Fire Inventory from NCAR (FINN) version 2.2 which provides daily varying global fire emissions at 1 x 1 km^2 resolution. FINN emissions are processed through SMOKE to enable inline plume rise of fire emissions within CMAQ. Biogenic emissions are calculated online within the model using the Biogenic Emission Inventory System (BEIS). Based on the input emissions, the CMAQ model simulates the concentrations, pollution transport, and distribution of the pollutants using several complex physical and chemistry equations by incorporating the impacts of meteorology (input from the UFS model) and physical processes. The chemical boundary conditions are based on 6-hourly Whole Atmosphere Community Climate Model (WACCM) simulations. The WACCM output is mapped onto CMAQ grids using the Initial Conditions Processor (ICON) and Boundary Conditions Processor (BCON). WACCM output provides the chemical initial conditions to CMAQ only on the 1st of June of every year. The chemical initial conditions for all the remaining days are set by recycling the chemical fields from the previous forecast cycle. To track the influence of wildfires on measurement stations, we have included a CO-FIRE tracer in the model which tracks the CO emitted by wildfires in the study domain and therefore is an indicator of the spread of the smoke from the wildfires across the CONUS. CO-FIRE is a chemically inert tracer that undergoes all the physical processes in the model as CO molecules do except for photo-chemistry. For each simulation day, we produced a 48-h forecast and saved hourly output for further analysis. The CMAQ output is collocated with AQS measurements using the "sitemap" utility of CMAQ.

3.2 The Analog Ensemble method

AnEn is a statistical-dynamical method that combines historical data and current predictions to generate an ensemble for increasing forecast accuracy. The method uses an archive of historical deterministic predictions paired with observations at those predictions' valid times to train the model. AnEn determines which historical forecasts are analogous to the current forecast by using a metric developed by Delle Monache et al. (2011). We employ two distinct sets of historical data: the first encompasses observed hourly $PM_{2.5}$ values collected from around 800 AQS sites over a span of 7 years (2015-2022). The second set comprises hourly forecasts generated by the CMAQ model, spanning the same 7-year duration and encompassing the identical locations as the AQS sites. Fundamentally, both datasets align in both time and space, with one providing observed historical data and the other offering forecasted values. Here, we use 10 analogous members. The best 10 analogs are chosen from each search of the archived datasets. The observations of $PM_{2.5}$ verifying these selected 10 analogous forecasts represent the 10

members of the $PM_{2.5}$ ensemble forecast. The optimal number of analogs (10 in this application) is based on a balance between sampling enough of the observed distribution while ensuring that all analogs are similar enough to the current prediction. In order to utilize AnEn for a forecast or estimation task, the initial requirement is to create a database comprising a sufficient number of entries (k) that adequately represent the phenomenon under consideration. The present-time model forecasts or analyses are subsequently compared to the elements in the database based on their resemblance to the features of the historical forecasts or analyses. This involves the identification of a set of predictors based on the model's output, and generating an analog by minimizing the given expression (Alessandrini et al. (2023)):

$$||C, H||_n = \sum_{i=1}^{N} \omega_i / \sigma_i \sqrt{(C_i - H_{in})^2}, \tag{1}$$

where C and H are the current and historical forecast analog of a predictor, N is the number of predictors, ω_i is the weight for each predictor, and σ_i is the standard deviation of that predictor in the historic data set. We used 6 years of data for the training process for every target year. We limit our target years to the years 2019, 2020, and 2021 since they had the highest history of large wildfires out of the 7 years. As already pointed out in (Alessandrini et al. (2019); Alessandrini (2022)), the AnEn introduces a positive bias when predicting the right tail of the forecast distribution. Here, the bias correction technique suggested by (Alessandrini (2022)) has been adopted to mitigate these biases. The forecasts in the distribution right tail are adjusted by adding a coefficient proportional to the difference between the target forecast and the mean of the analog forecasts.

3.3 Weight optimization

The AnEn approach is highly customizable, and several steps are involved in generating the weights, and analogs can vary depending on the specific application. When running AnEn without weight optimization, the algorithm assigns a weight (ω) of 1 to all predictors which means treating them as of equal importance. With the weight optimization on, the algorithm searches for the best possible weight combinations for predictors at every observation station in order to minimize the model RMSE in the forecast results at that specific station. As a result, the algorithm produces a file with a size of $M \times N$ (where M is the number of stations in the study and N is the number of predictors), which contains a set of optimal weights for each station and predictor. To accomplish this, the original version of the algorithm makes use of the so-called "brute force" approach, assigning ten possible weights to each predictor, i.e. 0.0-1.0 with 0.1 increments, with the constraints that they add up to 1, and calculated the RMSE at each scenario at every station (Delle Monache et al. (2020)). It then assigned a single combination of weights to the predictors at which the RMSE for that observing station was the lowest. However, by adding new predictors, there exists millions of combinations to be tested to be able to pick the best combination which requires high computational power and time. To resolve this issue, we followed a similar approach as Alessandrini et al. (2018). We first select the most important predictor (named P_1, $PM_{2.5}$ from CMAQ in this case) and compute the AnEn forecast based only on it and the forecast's performance as RMSE of the mean of the 10 members. Then, each of the remaining $NP - 1$ predictors, P_i, are tested one by one together with P_1. For each pair, AnEn predictions are generated with all the possible weight combinations, using a weight increment of 0.1 and the constraint (include sum equation to 1, see Alessandrini et al. (2018)). The pair resulting in the lowest RMSE determines

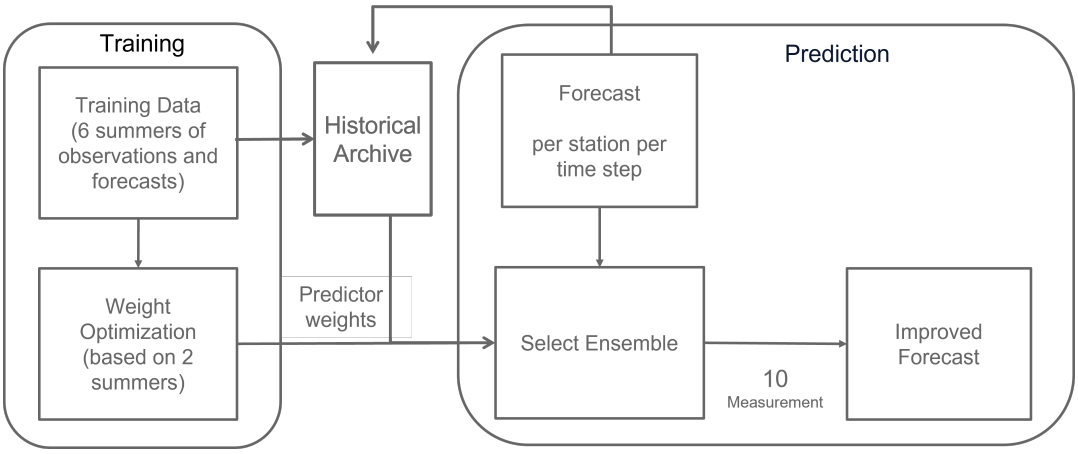

Figure 2. A schematic diagram of Analog Ensemble algorithm.

the second predictor, P_2, which is selected only if the improvement (decrease) of RMSE, compared to using P_1 alone, is more than 1%. If P_2 is chosen, the procedure is repeated to generate all possible triplets with the remaining NP - 2 predictors, along with P_1 and P_2. The procedure is interrupted when the increase in performance (decrease of RMSE), compared to the previous iteration, is lower than 1%. In this iterative procedure of weight selection, we added an additional constraint with respect to Alessandrini et al. (2018) to further limit the number of combinations to be tested. In fact, when testing a new predictor P_i, only $\omega_i < \omega_1 \dots \omega_{i-1}$ are tested. The selected set of predictors and their corresponding weights ω_i are used to generate the AnEn predictions over the verification dataset. The 1% threshold has been identified as an optimal choice to detect statistically significant improvements regarding RMSE. For weight optimization, we use a two-year-long (242 days) training period for each target year.

3.4 New Predictors

Six predictors (from the model) have been already used in the literature and operations, i.e., the CMAQ modeled $PM_{2.5}$, surface temperature, relative humidity, wind speed, wind direction, and the Planetary Boundary Layer (PBL) height (Delle Monache et al. (2020)). The importance of these meteorological predictors has been thoroughly studied in the literature Arya et al. (1999); Jacobson (2005). For instance, in a study by Li et al., the $PM_{2.5}$ concentrations showed negative correlations with temperature, relative humidity (RH), and wind speed. Furthermore, they confirm that wind direction plays a significant role in influencing $PM_{2.5}$ concentrations by determining the direction of dispersion. The findings of that study underscore the crucial impact of meteorological factors on the aggregation, diffusion, and spread of $PM_{2.5}$. These factors hold particular sway over $PM_{2.5}$ concentrations in instances where domestic emissions remained stable (Li et al. (2017)). A separate study by Wang and Ogawa (2015) confirms the negative correlation between $PM_{2.5}$ and temperature, as well as the critical role of the wind direction in pollution transport. However, they report that humidity and wind speed have threshold-dependent correlations with $PM_{2.5}$, with the direction (positive or negative) related to whether their values were below or above the threshold. The PBL height,

on the other hand, depends on atmospheric stability which is crucial in pollution transport Arya et al. (1999). Our selection of predictors is based on similar studies. In addition to the six predictors, we have developed and evaluated two new predictors to further enhance post-processing performance in events with elevated atmospheric $PM_{2.5}$ concentrations caused by wildfires. Specifically, the first new predictor is a modeled CO fire tracer ($CO - FIRE$) integrated into the CMAQ model, which tracks CO emissions from wildfires. The second new predictor is the maximum observed $PM_{2.5}$ concentration for the previous day at every station, regardless of forecast hour. For instance, for a given 48-hour forecast on June 2nd, we calculated the maximum observed $PM_{2.5}$ from AQS on June 1st and included this variable as a predictor in the AnEn algorithm. Our hypothesis is that, on days with highly polluted $PM_{2.5}$ concentrations, which may indicate the presence of a potential wildfire nearby, the maximum observed $PM_{2.5}$ from the previous day can indicate favorable analogous conditions for a fire event in the next 48 hours. Overall, the incorporation of these two new predictors is expected to improve the accuracy of our post-processing approach. The selection of predictors is strategic, enabling them not only to recognize past pollution episodes of comparable magnitude but also to identify the meteorological and chemical circumstances that contributed to previous air pollution incidents. To that end, since the focus of this project is wildfire events, we hypothesize that the two new predictors will help the algorithm detect the most analogous ensemble members and increase the accuracy of the forecasts.

3.5 Measurement Data

To obtain the necessary data for our study, we utilized on-site measurements from the Air Quality System (AQS) database (https://aqs.epa.gov/aqsweb/airdata/download_files.html#Raw). This database provides a comprehensive collection of air pollution, meteorological, and other relevant data from thousands of monitoring stations across the CONUS, and is operated by the EPA. We co-located the observed $PM_{2.5}$ concentrations and other meteorological factors from over 2300 sites within our study domain, and after careful analysis, we selected 795 stations for our study. Our selection process was based on data availability and quality, ensuring that each site contained valid data for at least 50% of the time from 2019 to 2021 (the target study years). This approach allowed us to develop a diverse and robust database with a reasonable range of variability which is necessary for AnEn post-processing. The distribution of the selected stations is presented in Fig. 1. To co-locate the measured $PM_{2.5}$ with the model, we extract the model data at the central mass point of every grid cell, which encompasses at least one of the AQS stations, at the exact hour that the measurement in that site has been made.

4 Results and Discussion

In this section, we will present the results of implementing AnEn on air quality forecasts generated by the CMAQ model. The results will be presented in four subsections. First, we will analyze the performance of the CMAQ model by evaluating its deterministic forecasts against on-site measurements from 795 observational sites located across the CONUS (see section 4.1). Once the shortcomings of the model outcomes and potential areas for improvement have been identified, we will assess the performance of the AnEn algorithm over the study period using observational data collected from the same 795 measurement sites (see section 3.5). We limit our major analysis to our target years 2019, 2020, and 2021, since they had the highest history

of large wildfires out of the 7 years of simulations. Additionally, we will present the results of the AnEn sensitivity analysis for the two new predictors introduced in this study in section 3.3. Finally, we will present the results of the weight optimization carried out by AnEn and the significance of each predictor in bias-correcting the CMAQ results in each EPA region (see section 4.2). We split our analysis to every geographic region defined by the EPA to understand the impacts of regional factors on model performance. We emphasize regions 8, 9, and 10, which experience high wildfire activity. These regions include the most wildfire-affected states such as California, Oregon, Colorado, and Washington.

4.1 UFS-CMAQ model performance

We studied the correlation between the model RMSE and $CO - FIRE$. The results show that the model RMSE for $PM_{2.5}$ is highly correlated with high episodes of $CO - FIRE$ (Fig. 3). This suggests that the model performs relatively poorly during wildfires.

Figure 4 shows the Mean Bias Error (MBE) calculated at every station with valid data over the years 2019-2021 and is scattered across the country to illustrate the model error, spatially. The results indicate that, in general, the CMAQ model shows lower MBE with a tendency for a slight systematically overestimating $PM_{2.5}$ concentrations at the eastern side of the CONUS in regions 2, 3, 4, and 5, with a few exceptions. Meanwhile, in western regions, the model shows overestimation and underestimation of $PM_{2.5}$ in the years 2019 and 2020-2021, respectively. The years 2020 and 2021 were important years for wildfire activities (https://www.nifc.gov/fire-information/statistics/wildfires) in the west and it is when we detect a significant underestimation of $PM_{2.5}$ in the western CONUS (Fig. 4). The underestimation of $PM_{2.5}$ in the west is an indicator that the CMAQ deterministic forecasts do not capture the high $PM_{2.5}$ episodes that could potentially be due to the wildfires in these regions. On the other hand, the slight overestimation on the east may be due to an overestimation of westerly wind speeds in the model which can result in an exaggerated transport of fire smoke towards the eastern regions of the country, creating an overestimation in the east and a more pronounced underestimation in the west. By referencing Figure 6b in the paper, in which we examine the performance of the UFS model, we can reasonably conclude that the model indeed overestimates wind speeds, which could be a reason for the overestimated $PM_{2.5}$ values on the east.

Amongst all EPA regions (see region definition here: https://www.epa.gov/aboutepa/regional-and-geographic-offices), the model performs the best in region 2 with the lowest RMSE value and it has the highest RMSE for regions 1, 9, and 10 (Fig. S2). This is important since our focus is mostly on the latter regions (9 and 10). In addition to regional analysis, we show the model MBE and RMSE calculated at every station versus the station longitude in figure 5a-b. It is clear that the model indicates significantly higher RMSE and a more negative MBE (an underestimation) in stations located on the western side of the CONUS. In terms of the error diurnal cycle, we find a systematic underestimation of $PM_{2.5}$ concentrations between hours 15:00-24:00 UTC and 39:00-48:00 (UTC) of the forecast time (daytime in all time zones) and a systematic overestimation between hours 05:00-15:00 UTC and 30:00-40:00 UTC (mostly nighttime in all regions) meaning that CMAQ bias follows a diurnal cycle as expected.

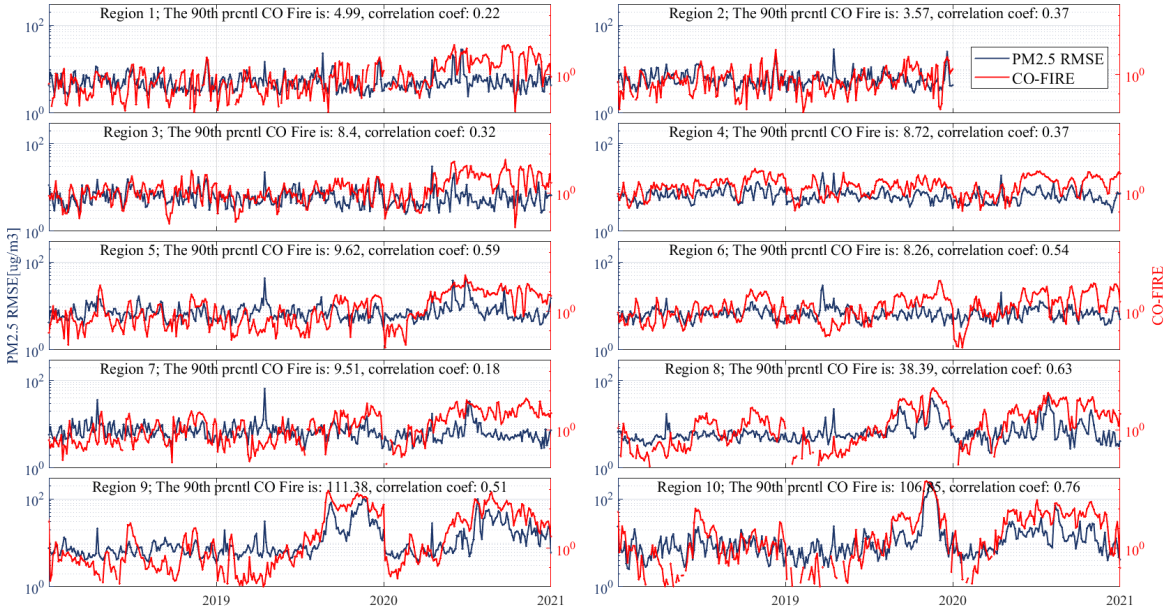

Figure 3. The relationship between the $PM_{2.5}$ RMSE from CMAQ forecasts (on the left axis in black), and $CO-FIRE$ (on the right axis in red). The Y axis is in logarithmic scale. The text on each plot shows the correlation coefficient between the RMSE and $CO-FIRE$ at every region when $CO-FIRE$ was higher than the 90th percentile value. The higher the correlation coefficient, the worse the model performance.

In addition to the CMAQ model, we evaluated the meteorological inputs created by the UFS model. Specifically, we analyzed the model performance for three variables - surface temperature, 10-m wind speed, and relative humidity - as we believe these variables play a critical role in determining $PM_{2.5}$ concentrations. Our results are presented in Fig. 6. We found that surface temperature and relative humidity exhibit strong correlations with the observations across all regions and sites, and at all forecast times (with the exception of region 7, where relative humidity is slightly overestimated). In contrast, the performance of the 10-m wind speed input showed a different pattern. We observed an overestimation of wind speed at almost all regions, except for region 10 where there was a slight underestimation. The overestimation in wind speed can eventually lead to an underestimation of $PM_{2.5}$ in those sites.

4.1.1 AnEn weight optimization

In this section, we discuss weight optimization by AnEn. In this regard, we highlight the significance of the eight predictors utilized in this study in enhancing the model outcomes. To evaluate the importance of each predictor, we analyze the statistical distribution of the weights generated by AnEn in ten EPA regions. This allows us to examine the relevance of each predictor in regions with specific characteristics. We are particularly interested in examining the impact of the two new predictors on correcting CMAQ forecasts in regions with a high incidence of wildfires. Our findings are presented in Fig. 4.1.1, where we observe that the $PM_{2.5}$ predictor (from the model forecast) has the greatest impact, as anticipated, in all regions. Interestingly,

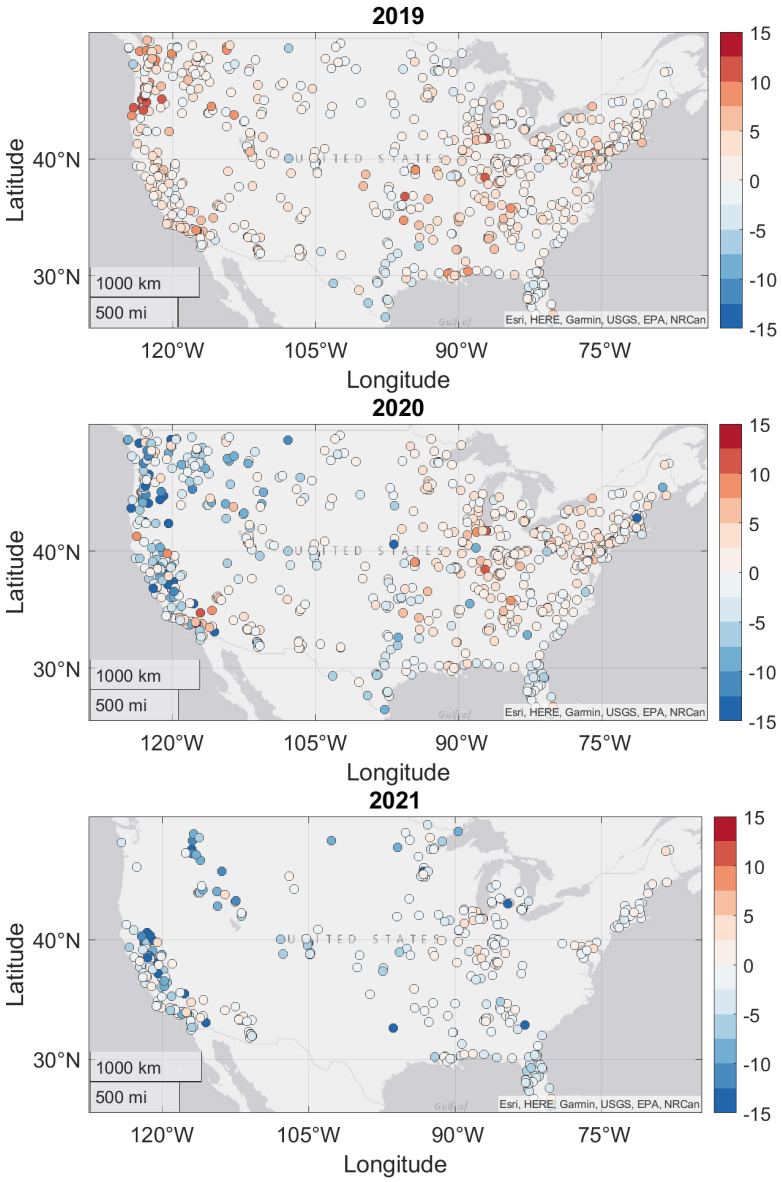

Figure 4. Mean Bias Error for forecasted $PM_{2.5}$ from the CMAQ model, calculated at every station across forecast hours (25-48) and all simulation days in every year. Blue shades represent an underestimation while red shades represent an overestimation in predictions.

in addition to $PM_{2.5}$, both $CO-FIRE$ and $PM_{2.5}-preday-max$ have the two highest weights in regions 7, 8, 9, and 10, making them two of the most critical predictors employed in our bias correction. The role of surface temperature in predicting $PM_{2.5}$ concentrations is particularly significant in regions 1 and 2 where wildfires are scarce. This finding is consistent with expectations. However, our analysis indicates that, among all the predictors considered, wind speed and PBL height tend to have

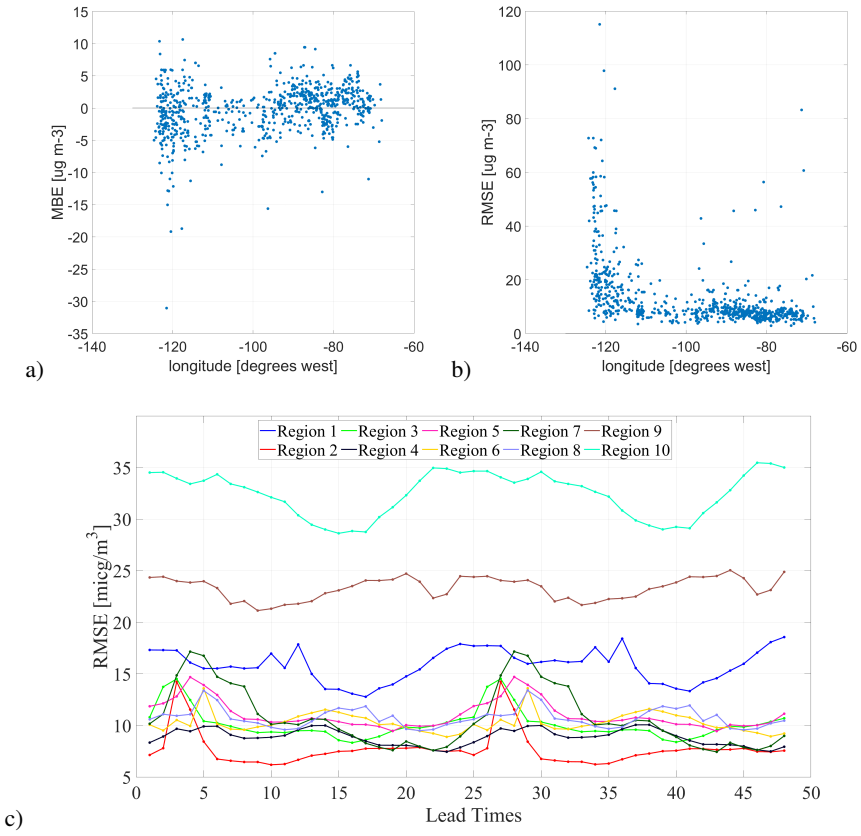

Figure 5. CMAQ model performance for $PM_{2.5}$; a) calculated MBE at every station scattered based on the station longitude, b) same as (a) but for the RMSE, and c) the RMSE calculated in ten EPA regions where each line represents the RMSE calculated across all days and stations in that specific region.

the lowest weights across most regions. Future studies may choose to prioritize the use of predictors that have a greater impact on their study parameters. In our analysis, we decided to include all 8 predictors, as we believe that each one of them could potentially affect $PM_{2.5}$ concentrations in the atmosphere. The selection of predictors is strategic in that they not only enable the identification of past pollution episodes of similar magnitude but also facilitate the identification of the meteorological and chemical conditions that have led to such episodes in the past.

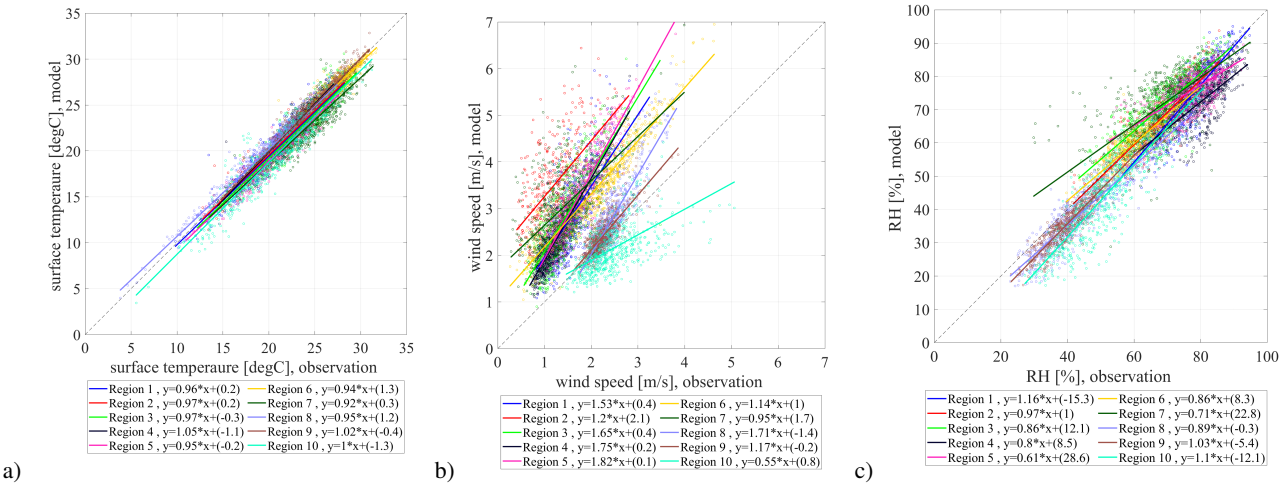

Figure 6. Meteorological input evaluation for the UFS model for a) surface temperature, b) 10m wind speed, and c) relative humidity. Different colors illustrate different EPA regions.

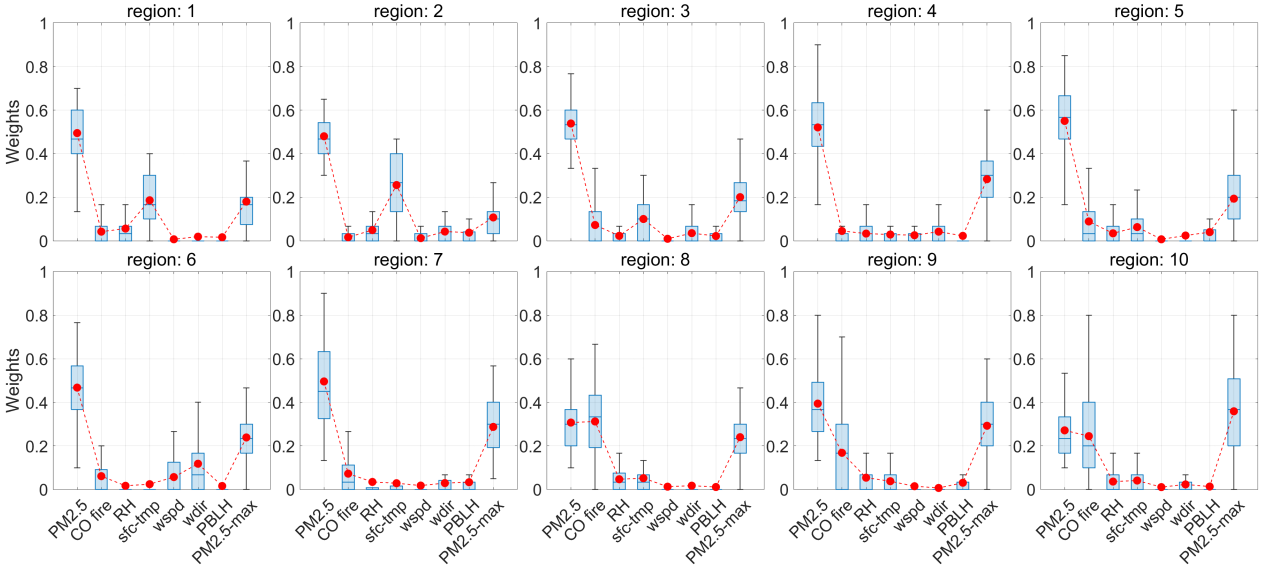

Figure 7. Weights assigned to each predictor by AnEn weight optimization algorithm, for each EPA region. The EPA regions have been defined in US. EPA (2023)

4.1.2 AnEn predictors

In this section, we assess the impact of the two novel predictors, $CO-FIRE$ and $PM_{2.5}-preday-max$, on the performance of the AnEn algorithm. We observe that both of the new predictors, $CO-FIRE$ and $PM_{2.5}-preday-max$, have a significant positive effect on the AnEn's performance. To evaluate their importance, we first applied AnEn on CMAQ outputs with the six predictors used in previous studies (referred to as "AnEn-6prdctr"). We then introduced $CO-FIRE$ and $PM_{2.5}-preday-max$ as the 7th and 8th predictors, respectively, and tested the performance of the algorithm in two additional scenarios referred to as "AnEn-7prdctr" and "AnE-8prdctr". To be consistent with the main AnEn run ("AnEn-8prdctr"), the sensitivity of the results to both predictors was assessed by using weight optimization in all test runs. In terms of overall RMSE, which was calculated over all stations and days in the study period at every forecast hour, the raw CMAQ forecasts showed the highest RMSE compared to any other scenario that included AnEn-corrected results. AnEn with traditional 6 predictors decreased the RMSE by up to 8% at maximum. When the "AnEn-7prdctr" scenario was run, which included $CO-FIRE$ as the new predictor, the RMSE decreased by another 6-8%, resulting in a 12% reduction in RMSE compared to the raw CMAQ forecasts. This indicates that the newly added $CO-FIRE$ predictor decreased the overall RMSE and improved AnEn performance across all forecast hours. The final run used in this study, "AnEn-8prdctr", included the $PM_{2.5}-preday-max$ predictor as the 8th predictor in the AnEn post-processing. The results showed that the $PM_{2.5}-preday-max$ predictor decreased the RMSE by an additional 15%, at maximum (Fig. 8). The results indicated that AnEn with 8 predictors, including the newly added predictors, consistently and significantly lowered the overall RMSE of the CMAQ forecasts, reducing it by up to 4.5 $\mu g/m^3$ (25%). The correlation coefficient follows a similar trend to the RMSE, with CMAQ exhibiting the lowest correlation with the observations. However, both "AnEn-6prdctr" and "AnEn-7prdctr" improved the correlation coefficient, with increases of up to 12% and 22%, respectively. The "AnEn-8prdctr" scenario demonstrated the highest correlation coefficient, with an increase of up to 40%. As for the MBE, "AnEn-6prdctr" and "AnEn-7prdctr" show comparable results to "AnEn-8prdctr," although "AnEn-7prdctr" has a slightly higher underestimation of 0.03 $\mu g/m^3$.

4.2 AnEn implementation

We use 8 predictors, including two new predictors, and a six-summer period of training time (726 days) for each study year to create AnEn corrected forecasts. We have found that AnEn outperforms the deterministic raw forecasts of the CMAQ model with a lower RMSE in all regions. As shown in Fig. 10a AnEn decreased the RMSE by up to 25% at every lead time. We present the relationship between the daily averaged $PM_{2.5}$ observations and $PM_{2.5}$ forecasts from CMAQ and AnEn in figure 9, where we find a significant increase in R^2 with AnEn corrected forecasts ($R^2 = 0.8$) compared to CMAQ raw forecasts ($R^2 = 0.2$). In addition, in Figures S1-S4, we provide a separate analysis of AnEn's performance for each EPA region. In this study, we will denote special attention to regions 8, 9, and 10, which have a higher history of wildfires. The largest decrease in RMSE due to AnEn was observed in regions 9 and 10 (as shown in Fig. S1, and Fig. S2) during the months of August and September, which are typically associated with higher wildfire activity. Conversely, the lowest reduction in RMSE was observed in June. The AnEn bias is more dampened compared to the bias in CMAQ (Fig. 10b). CMAQ bias shows a diurnal cycle, while AnEn

eliminates diurnal error variations because AnEn uses observations corresponding to each hour of a day to correct CMAQ forecasts at each lead time independently. Consequently, the AnEn bias is more consistent and shows an overall underestimation of approximately 0.5 $\mu g/m^3$ while CMAQ MBE varies between -1 to +1.2 $\mu g/m^3$. It is evident that the correlation coefficient between AnEn and AQS is consistently higher than that between CMAQ and AQS at all forecast times (20%-40% higher correlation depending on forecast time), as shown in Fig. 10c. Figures 8 and 10 provide a comprehensive overview of the model's performance, illustrating a recurring diurnal pattern in the model RMSE across the entire domain, containing four different time zones. The plots show that the model consistently performs at its best from 10:00 to 20:00 UTC, indicated by the lower RMSE (Figure 8) and MBE (Figure 10) values on each forecast day. Despite the presence of four distinct time zones within the domain, the periods of low RMSE on these charts consistently correspond to the early morning and daytime hours across all regions and time zones. On the other hand, the hours with higher RMSE and MBE values consistently align with nighttime hours in all four time zones. This consistent pattern suggests that the model has a better performance during daytime hours while displaying a poor performance during nighttime hours when predicting PM2.5 values. The three statistical metrics used, namely the RMSE, MBE, and correlation coefficient, all provide clear evidence that AnEn significantly improves the accuracy of $PM_{2.5}$ forecasts.

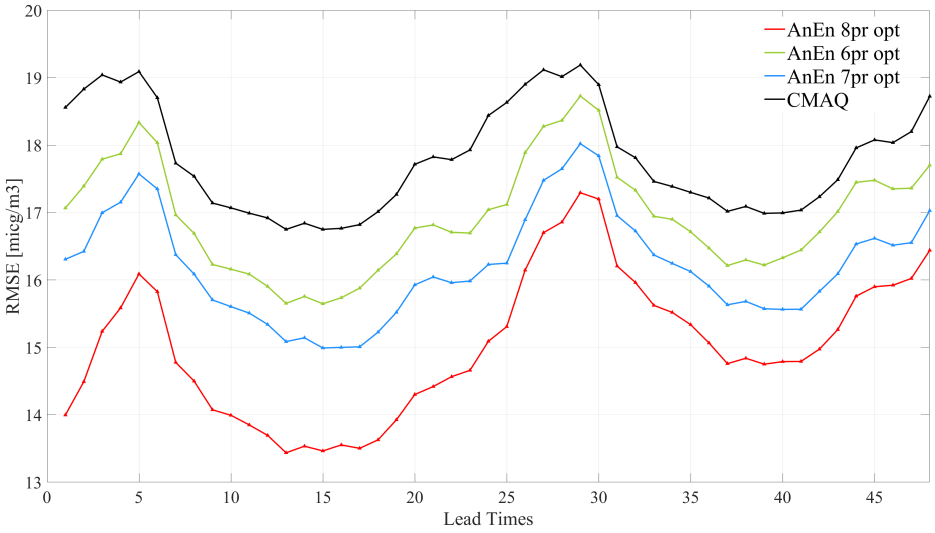

Figure 8. $PM_{2.5}$ RMSE vs. forecast lead time for CMAQ (black), AnEn with previously used 6 predictors (green), AnEn with 7 predictors including new $CO-FIRE$ predictor (blue), and AnEn with 8 predictors including $PM_{2.5}-preday-max$ (red). Calculations are averages over all sites during the periods of study described in the text.

The concentrations of the $PM_{2.5}$ forecasts by CMAQ and AnEn are compared to observations in Fig. 11a, where, we detect a better match between AQS and AnEn especially in high $PM_{2.5}$ bins. Figure 11b depicts the time series of daily RMSE (red) and MBE (blue) calculated in the study period, with the left and right axes representing the respective values. Notably, the RMSE

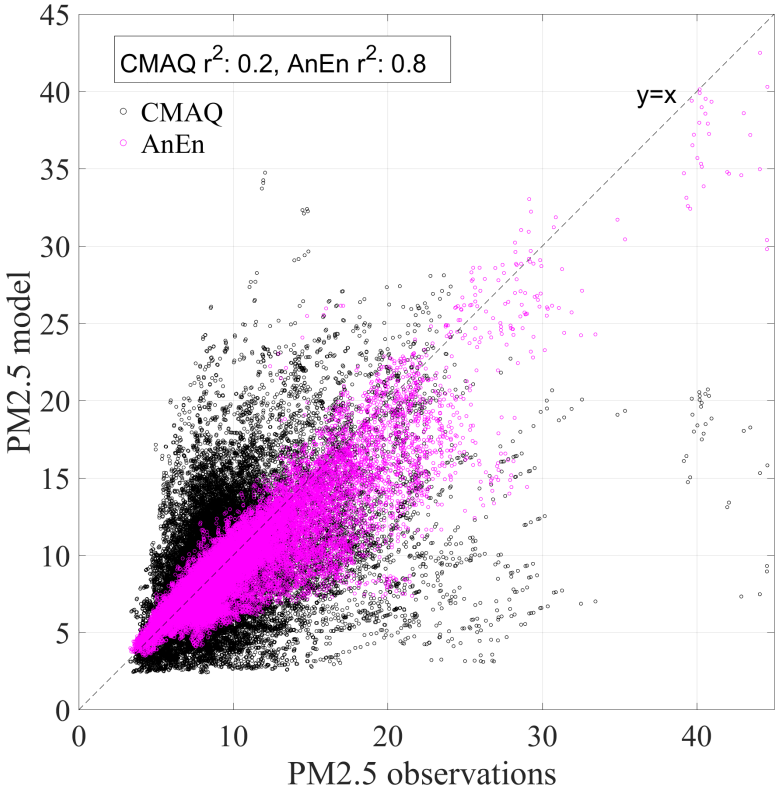

Figure 9. $PM_{2.5}$ observations vs $PM_{2.5}$ forecasts; black dots represent the CMAQ deterministic forecasts and pink dots are the AnEn corrected forecasts. The dots represent daily averaged data at every forecast hour and observing site.

has decreased significantly for all days in the simulation period using AnEn. In terms of MBE, AnEn performs better overall, although a few days exhibit a slight increase. The shaded areas represent the error range between the 33rd and 67th percentile of the error data, and the figure highlights that this range is narrower with AnEn than with CMAQ raw forecasts. Figure 11 represents the AnEn performance based on all stations across the CONUS. It provides a general look into AnEn's performance regardless of the location. However, we expand this analysis later in 12, by zooming into three individual wildfire cases in three stations located in California (shown in blue in Figure 1a). We study these stations during the fire period (from start to finish), specific to each station.

In Figure S5, we present the $PM_{2.5}$ levels only on days when $PM_{2.5}$ observed levels exceeded the 80th percentile. We compare $PM_{2.5}$ levels from AQS measurements, CMAQ forecasts, and forecasts after AnEn post-processing. Our findings indicate that the AnEn-corrected $PM_{2.5}$ values during potential wildfire days are more consistent with the measured values. While AnEn and AQS match best during days with lower $PM_{2.5}$ concentrations, AnEn still significantly improves the $PM_{2.5}$ forecast results during the highest $PM_{2.5}$ episodes. To further analyze this, we created a contingency table (2) to assess the model's ability to predict $PM_{2.5}$ during wildfires in regions 8, 9, and 10. We assumed that events with observed $PM_{2.5}$ concentrations exceeding

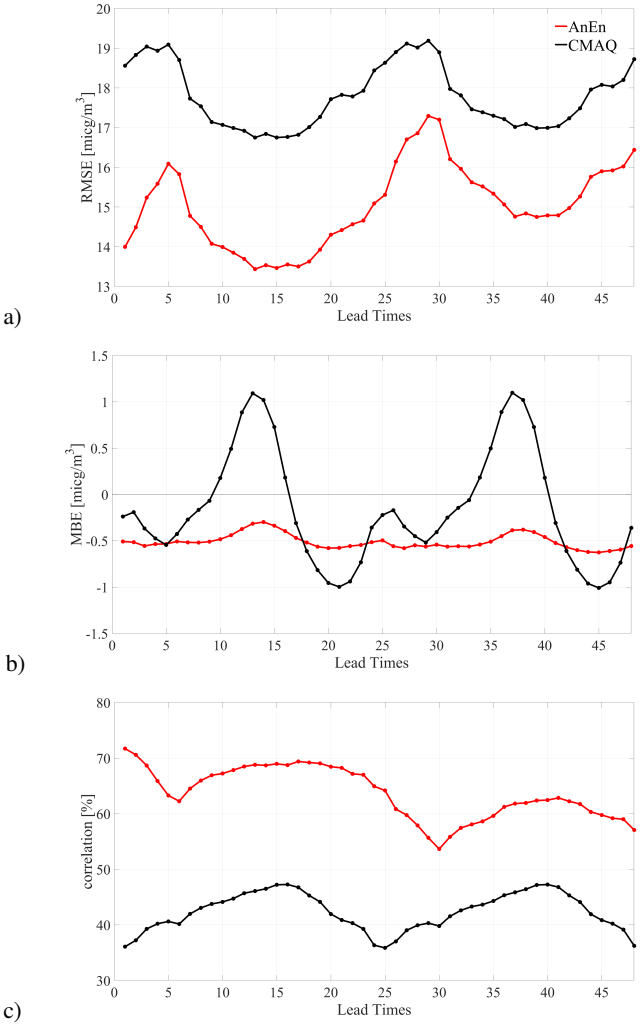

Figure 10. RMSE (top), bias (middle), and correlation (bottom), of $PM_{2.5}$ vs. lead time in forecasts from CMAQ (black), AnEn mean (red). Calculations are averages over all sites during the periods of study described in the text.

100 $\mu g/m^3$ represent wildfire incidents. Fig. 5 summarizes these results. In the contingency table, a "Hit" occurs when both observed and model events are "YES," indicating that the model predicted the event correctly. If the event-observed and model are both 'NO', it is a 'correct non-event' meaning that the model predicted correctly that it is not a high $PM_{2.5}$ day. On the other hand, if the event observed is 'YES' and the model is 'NO', it is a 'Miss', suggesting that the model failed to predict the event. Similarly, if the observed event is "NO," and the model is "YES," it is a "False Alarm," suggesting that the model falsely predicted a high $PM_{2.5}$ episode.

Figure 2 shows that AnEn outperformed CMAQ in detecting high $PM_{2.5}$ episodes during wildfires, as indicated by significantly higher numbers of hits in all three regions (more than doubled hits). Additionally, AnEn showed lower numbers of missed events than CMAQ in all three regions. However, CMAQ demonstrated a better performance in avoiding false alarms, with lower rates of "False Alarms" compared to AnEn in all three regions. This can be due to the fact that, in a decreasing $PM_{2.5}$ trend after a high episode, AnEn takes a day or two to be trained and capture the trend and therefore reports higher values than the observations. The Critical Success Index (CSI) is calculated for CMAQ and AnEn separately at each region. The CSI is a verification measure of categorical forecast performance, and it is equal to the total number of hits divided by the total number of storm forecasts plus the number of misses (hits + false alarms + misses) (Wilks (2011)). The CSI indicates that the AnEn is outperforming CMAQ in all three regions by more than the double CSI.

Table 2. Contingency table; green cells indicate the number of times that the model (CMAQ or AnEn) predicted an event correctly by either detecting that PM2.5 was higher or lower than 100 g/m3 (here, considered as a threshold for wildfire indication). The red cells indicate the number of times that the model either missed a high PM2.5 event or predicted a high PM2.5 event falsely.

			Event Observed		
			YES	NO	CSI
Region 8	CMAQ	YES	93	65	0.09
		NO	855	377505	
	AnEn	YES	219	113	0.2
		NO	729	377457	
Region 9	CMAQ	YES	815	1121	0.07
		NO	8735	918273	
	AnEn	YES	1854	1839	0.16
		NO	7696	917553	
Region 10	CMAQ	YES	1534	320	0.11
		NO	11337	660341	
	AnEn	YES	3760	1371	0.26
		NO	9112	659290	

We have utilized high episodes of $PM_{2.5}$ observations as well as $CO-FIRE$ as indicators of potential wildfires in our study. For most regions, AnEn corrected $PM_{2.5}$ values match the observations very well, while there is a clear under/overestimation of the $PM_{2.5}$ values by CMAQ based on the region. However, in western regions, the results are not as close to observations as in other regions. This is due to the challenges associated with predicting $PM_{2.5}$ levels in areas affected by wildfires, which can result in large and sudden increases in $PM_{2.5}$ levels. Next, we will take a closer look at the stations in western CONUS that were impacted by the wildfires during the study period.

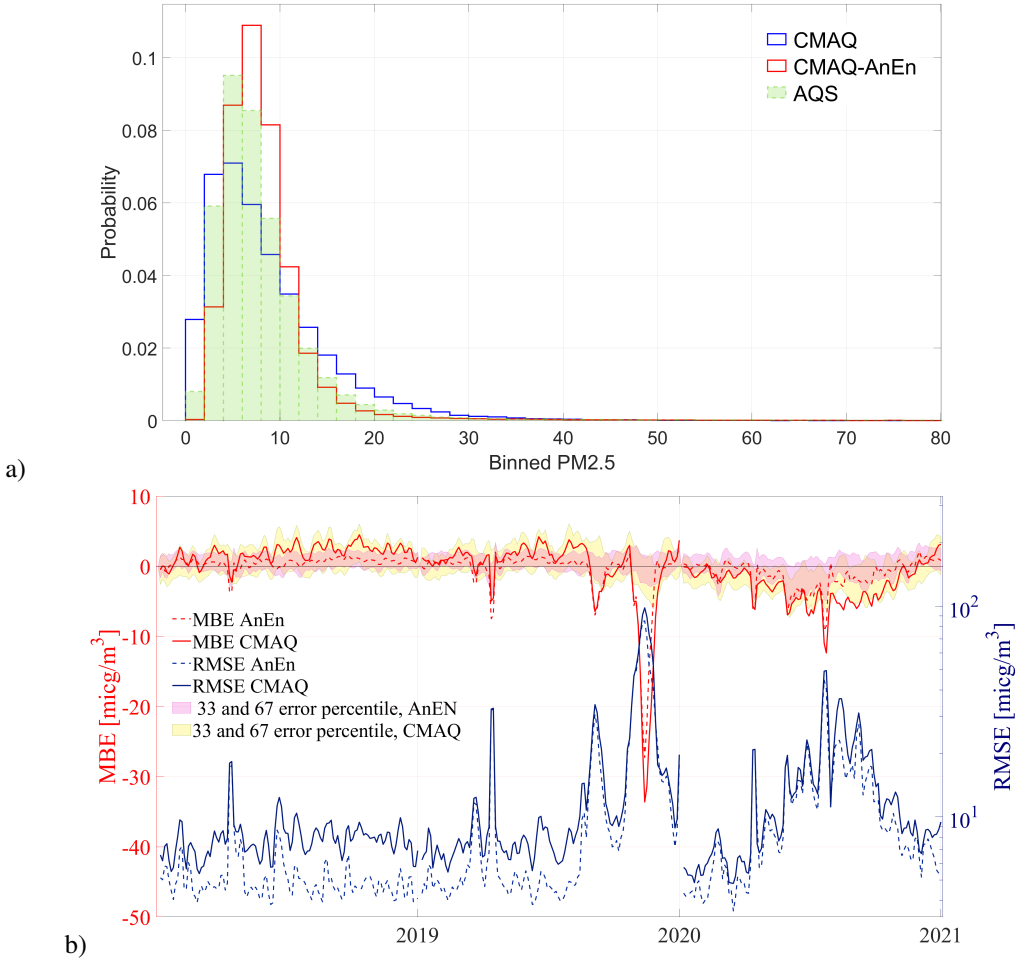

Figure 11. a) $PM_{2.5}$ concentrations for AQS, CMAQ, and AnEn in green, blue, and red stairs, respectively. b) The RMSE and MBE time series. The left axis in red is MBE and the right axis (logarithmic scale) in blue is RMSE for CMAQ in solid lines and AnEn in dashed lines. The shaded area is the error range between the 33rd and 67th percentile (centered at 0). AnEn (in the pink shade) has a narrower range compared to CMAQ.

Case Study:

We analyzed the AnEn performance in three stations with the highest average concentrations of $PM_{2.5}$ and $CO-FIRE$ over time, assuming they represent locations of significant wildfires during the study period. These stations are located in California, as depicted in blue in Fig. 1. $PM_{2.5}$ concentrations for each day at each of the three stations are illustrated in Fig. 12, which zooms in on days with the highest observed $PM_{2.5}$ concentrations at each station, confirmed to be representative of wildfires. The $PM_{2.5}$ values are the averages over the forecast hours for each day. The AnEn algorithm improved the forecast accuracy at these stations during wildfires. We observed a consistent pattern in AnEn performance across all three stations: the RMSE decreased by up to 300 $\mu g/m^3$ (in station one, not shown) during peak concentrations. It is clear in Fig. 12 that the AnEn predicted values are closer to the observed concentrations in the peaks. Initially, the AnEn showed a similar pattern to the CMAQ model when $PM_{2.5}$ began an increasing trend. However, after approximately one day, the AnEn algorithm caught up and significantly reduced the RMSE, indicating an increase in forecast accuracy. MBE exhibited a similar pattern, with AnEn reducing both MBE and RMSE during the highest episodes of $PM_{2.5}$. Nevertheless, on some days, the MBE was higher than the CMAQ forecasts (not shown).

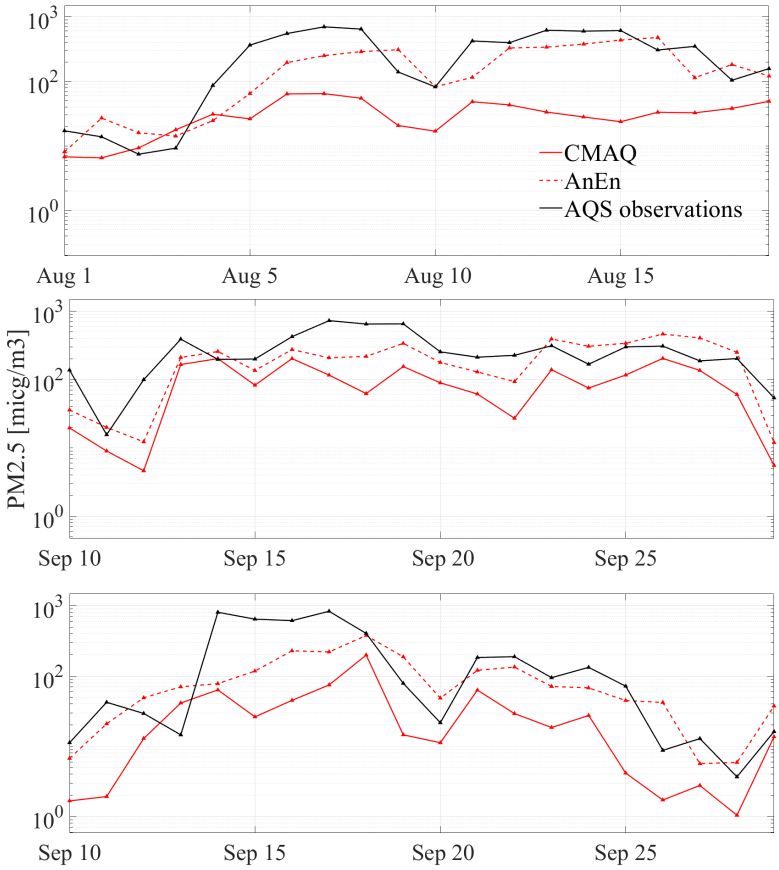

Figure 12. $PM_{2.5}$ concentrations averaged over forecast lead times in three selected stations for AQS, CMAQ model), and AnEn in solid black, solid red, and dashed red lines, respectively. These three stations had the highest average $PM_{2.5}$ levels in the entire domain and all three stations are located in California (shown in blue dots in Fig. 1)

5 Conclusions

Air quality forecasts provide valuable information for the control of air pollution; however, the accuracy of the forecasts is only sometimes favorable, and it is extremely challenging to produce good forecasts of pollutants during extreme events such as wildfires. In this study, we created 48-hour PM2.5 forecasts over the CONUS for seven consecutive major fire seasons (June 1st to Sep 29th) during 2015-2021 using the UFS-CMAQ modeling system with a resolution of 12 x 12 km2 with 442, 265 grid cells horizontally, and 35 grid points, vertically. We included a $CO-FIRE$ tracer in the model to track the CO emitted by wildfires in our domain and we co-located the CMAQ outputs with the measurements from the AQS from 795 stations across our domain. Our research body is divided into three main parts. First, we assess the accuracy of the deterministic forecasts produced by the CMAQ model. Next, we contribute to the field by pioneering the application of the state-of-the-art Analog Ensemble,

a statistical-dynamical post-processing method, to explore potential improvements to forecast accuracy during wildfires while incorporating a tracer to precisely trace the fire smoke, and eventually and in parallel to the second objective, for the first time, we introduce two novel predictors and we test their ability and importance in enhancing the forecast accuracy as is the ultimate goal of this study. Our investigation focused on $PM_{2.5}$ concentrations since it is one of the most related pollutants to wildfires. The two new predictors are $CO-FIRE$ from the model and $PM_{2.5}-preday-max$, which is the maximum observed $PM_{2.5}$ value from the previous day and is obtained from the observations. We devote special attention to the western part of CONUS, including EPA regions 8, 9, and 10 which experience the highest wildfire activity.

Upon evaluating the CMAQ deterministic forecasts with AQS observations, we found a high bias in the model outputs, especially a significant underestimation in the western CONUS during the years with active wildfires. This means that the model did not properly capture the increase in $PM_{2.5}$ values during the extreme events. We post-processed the data with AnEn using 8 predictors, including the six predictors that are already used in the operations, plus two new predictors introduced here, and a six-summer period of training period (726 days) for each study year and weight optimization by AnEn, which assigns the optimal weights to the defined predictors.

We find that AnEn increases model accuracy by decreasing the RMSE, increasing the forecast correlation with observations, and increasing/decreasing the model bias in different lead times. Similar to the CMAQ deterministic forecasts, AnEn matches best with AQS data during days with low $PM_{2.5}$ concentrations (with significant improvement over CMAQ) and significantly improves the $PM_{2.5}$ forecast results during the highest $PM_{2.5}$ episodes. The model RMSE decreased by up to 25% when considering all stations across the domain. In stations with wildfire events, the RMSE decreased by up to 300 $\mu g/m^3$. In addition, the correlation between AnEn forecasts and observations was 20% - 40% higher than that between CMAQ and observations. When looking at the $PM_{2.5}$ trend during wildfires, AnEn performs similarly to CMAQ in the initial phase of a fire event, but it soon catches up and decreases the error significantly. In addition, we find that the two new predictors introduced in this study ($CO-FIRE$ and $PM_{2.5}-preday-max$), along with the predictor $PM_{2.5}-model$ (obtained from the model) have the highest weights in correcting the model outputs, especially in regions with high wildfire risks.

This pioneering effort provides valuable insights to the atmospheric science community and will guide future research endeavors in this domain. Nonetheless, there is still scope for future studies to improve the forecast accuracy further. In a forthcoming study, we will expand upon this research by incorporating advanced techniques to interpolate observations across the entire domain, rather than solely relying on the stations utilized in this current study. This extended investigation will encompass the application of AnEn to the complete domain, beyond just the observational station locations. Additionally, it will explore an in-depth examination of various techniques for data interpolation and assess their impact on the performance of AnEn.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements. This manuscript was prepared by all NCAR authors using Federal funds under award NA19OAR4590083 from the NOAA Office of Oceanic and Atmospheric Research, U.S. Department of Commerce. The statements, findings, conclusions, and recommendations are those of the author(s) and do not necessarily reflect the views of the NOAA Office of Oceanic and Atmospheric Research or the U.S. Department of Commerce. The NACC model used in this work was developed under funding from UMD/NOAA/CISESS: GMU Air Surface Exchange and Atmospheric Composition Research (Sponsor Number: 79785-Z7554202 Amend A). We would like to acknowledge high-performance computing support from Cheyenne (doi: 10.5065/D6RX99HX (accessed on 31 May 2023)), provided by NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation. The National Center for Atmospheric Research is sponsored by the National Science Foundation under Cooperative Agreement 1852977.

References

Alessandrini, S., 2022. Predicting rare events of solar power production with the analog ensemble. Solar Energy 231, 72–77.

Alessandrini, S., Delle Monache, L., Rozoff, C.M., Lewis, W.E., 2018. Probabilistic prediction of tropical cyclone intensity with an analog ensemble. Monthly Weather Review 146, 1723–1744.

Alessandrini, S., Delle Monache, L., Sperati, S., Cervone, G., 2015. An analog ensemble for short-term probabilistic solar power forecast. Applied energy 157, 95–110.

Alessandrini, S., Kim, J.H., Jimenez, P.A., Dudhia, J., Yang, J., Sengupta, M., 2023. A gridded solar irradiance ensemble prediction system based on wrf-solar eps and the analog ensemble. Atmosphere 14, 567.

Alessandrini, S., Sperati, S., Delle Monache, L., 2019. Improving the analog ensemble wind speed forecasts for rare events. Monthly Weather Review 147, 2677–2692.

Ancell, B.C., Bogusz, A., Lauridsen, M.J., Nauert, C.J., 2018. Seeding chaos: The dire consequences of numerical noise in NWP perturbation experiments. Bulletin of the American Meteorological Society 99, 615 – 628. doi:https://doi.org/10.1175/BAMS-D-17-0129.1.

Arya, S.P., et al., 1999. Air pollution meteorology and dispersion. volume 310. Oxford University Press New York.

Buizza, R., 2008. The value of probabilistic prediction. Atmospheric Science Letters 9, 36–42.

Buizza, R., Milleer, M., Palmer, T.N., 1999. Stochastic representation of model uncertainties in the ecmwf ensemble prediction system. Quarterly Journal of the Royal Meteorological Society 125, 2887–2908.

Campbell, P.C., Tang, Y., Lee, P., Baker, B., Tong, D., Saylor, R., Stein, A., Huang, J., Huang, H.C., Strobach, E., et al., 2022. Development and evaluation of an advanced national air quality forecasting capability using the noaa global forecast system version 16. Geoscientific Model Development 15, 3281–3313.

Delle Monache, L., Alessandrini, S., Djalalova, I., Wilczak, J., Knievel, J.C., Kumar, R., 2020. Improving air quality predictions over the united states with an analog ensemble. Weather and Forecasting 35, 2145–2162.

Delle Monache, L., Eckel, F.A., Rife, D.L., Nagarajan, B., Searight, K., 2013. Probabilistic weather prediction with an analog ensemble. Monthly Weather Review 141, 3498–3516.

Delle Monache, L., Nipen, T., Liu, Y., Roux, G., Stull, R., 2011. Kalman filter and analog schemes to postprocess numerical weather predictions. Monthly Weather Review 139, 3554–3570.

Djalalova, I., Wilczak, J., McKeen, S., Grell, G., Peckham, S., Pagowski, M., DelleMonache, L., McQueen, J., Tang, Y., Lee, P., et al., 2010. Ensemble and bias-correction techniques for air quality model forecasts of surface o3 and pm2. 5 during the texaqs-ii experiment of 2006. Atmospheric Environment 44, 455–467.

Du, J., Mullen, S.L., Sanders, F., 1997. Short-range ensemble forecasting of quantitative precipitation. Monthly Weather Review 125, 2427–2459.

Ebert, E.E., 2001. Ability of a poor man's ensemble to predict the probability and distribution of precipitation. Monthly Weather Review 129, 2461–2480.

Foley, K.M., Hogrefe, C., Pouliot, G., Possiel, N., Roselle, S.J., Simon, H., Timin, B., 2015. Dynamic evaluation of cmaq part i: Separating the effects of changing emissions and changing meteorology on ozone levels between 2002 and 2005 in the eastern us. Atmospheric Environment 103, 247–255.

Galmarini, S., Bianconi, R., Bellasio, R., Graziani, G., 2001. Forecasting the consequences of accidental releases of radionuclides in the atmosphere from ensemble dispersion modeling. Journal of Environmental Radioactivity 57, 203–219.

Golbazi, M., Archer, C.L., Alessandrini, S., 2022. Surface impacts of large offshore wind farms. Environmental Research Letters 17, 064021.

Hamill, T.M., Whitaker, J.S., 2006. Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. Monthly Weather Review 134, 3209–3229.

Jacobson, M.Z., 2005. Fundamentals of atmospheric modeling. 2 ed., Cambridge University Press.

Kumar, R., Lee, J.A., Delle Monache, L., Alessandrini, S., 2019. Effect of meteorological variability on fine particulate matter simulations over the contiguous united states. Journal of Geophysical Research: Atmospheres 124, 5669–5694.

Lewis, W.E., Olander, T.L., Velden, C.S., Rozoff, C., Alessandrini, S., 2021. Analog ensemble methods for improving satellite-based intensity estimates of tropical cyclones. Atmosphere 12, 830.

Li, X., Feng, Y., Liang, H., 2017. The impact of meteorological factors on pm2. 5 variations in hong kong, in: IOP Conference Series: Earth and Environmental Science, IOP Publishing. p. 012003.

McKeen, S., Chung, S., Wilczak, J., Grell, G., Djalalova, I., Peckham, S., Gong, W., Bouchet, V., Moffet, R., Tang, Y., et al., 2007. Evaluation of several pm2. 5 forecast models using data collected during the icartt/neaqs 2004 field study 112.

Molteni, F., Buizza, R., Palmer, T.N., Petroliagis, T., 1996. The ecmwf ensemble prediction system: Methodology and validation. Quarterly journal of the royal meteorological society 122, 73–119.

Palmer, T.N., 2002. The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography 128, 747–774.

Ryu, Y.H., Hodzic, A., Barre, J., Descombes, G., Minnis, P., 2018. Quantifying errors in surface ozone predictions associated with clouds over the conus: a wrf-chem modeling study using<? xmltex\break?> satellite cloud retrievals. Atmospheric Chemistry and Physics 18, 7509–7525.

Stensrud, D.J., Bao, J.W., Warner, T.T., 2000. Using initial condition and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. Monthly Weather Review 128, 2077–2107.

Toth, Z., Kalnay, E., 1993. Ensemble forecasting at nmc: The generation of perturbations. Bulletin of the american meteorological society 74, 2317–2330.

U.S. Environmental Protection Agency (EPA), 2020. Particulate Matter (PM) pollution. https://www.epa.gov/pm-pollution/particulate-matter-pm-basics#PM, accessed 06/10/2020.

U.S. Environmental Protection Agency (EPA), 2022a. NAAQS Table. https://www.epa.gov/criteria-air-pollutants/naaqs-table, accessed 04/04/2022.

U.S. Environmental Protection Agency (EPA), 2022b. Photochemical Air Quality Modeling. Support Center for Regulatory Atmospheric Modeling (SCRAM) URL: https://www.epa.gov/scram/photochemical-air-quality-modeling. accessed on 12/01/2022.

US. EPA, 2023. Regional and geographic offices. https://www.epa.gov/aboutepa/regional-and-geographic-offices.

Wang, J., Ogawa, S., 2015. Effects of meteorological conditions on pm2. 5 concentrations in nagasaki, japan. International journal of environmental research and public health 12, 9089–9101.

Weigel, A.P., Liniger, M., Appenzeller, C., 2008. Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography 134, 241–260.

Wilks, D.S., 2011. Statistical methods in the atmospheric sciences. volume 100. Academic press.

Zemouri, N., Bouzgou, H., Gueymard, C.A., 2019. Multimodel ensemble approach for hourly global solar irradiation forecasting. The European Physical Journal Plus 134, 594.

Zhang, F., Bei, N., Nielsen-Gammon, J.W., Li, G., Zhang, R., Stuart, A., Aksoy, A., 2007. Impacts of meteorological uncertainties on ozone pollution predictability estimated through meteorological and photochemical ensemble forecasts. Journal of Geophysical Research: Atmospheres 112.

Ziehmann, C., 2000. Comparison of a single-model eps with a multi-model ensemble consisting of a few operational models. Tellus A 52, 280–299.

SUPPLEMENTARY INFORMATION

This appendix contains additional figures from the study domains as well as some definitions that are used in our methods section.

Correlation Coefficient:

$$(X,Y) = cov(X,Y)/[var(X) \times var(Y)]^{1/2}$$

Error :

$E = A_i - O_i$,

where A is the AnEn outputs and O is the observations.

RMSE :

$$RMSE = 1/N \sum i = 1N(A_i - O_i)^2$$

MBE :

$$MBE = 1/N \sum i = 1N(A_i - O_i)$$

STD :

$$STD = 1/(N-1)i = 1N|Ai - \mu|2 \ ,$$

Where,

$$\mu = 1/Ni = 1N \sum Ai$$

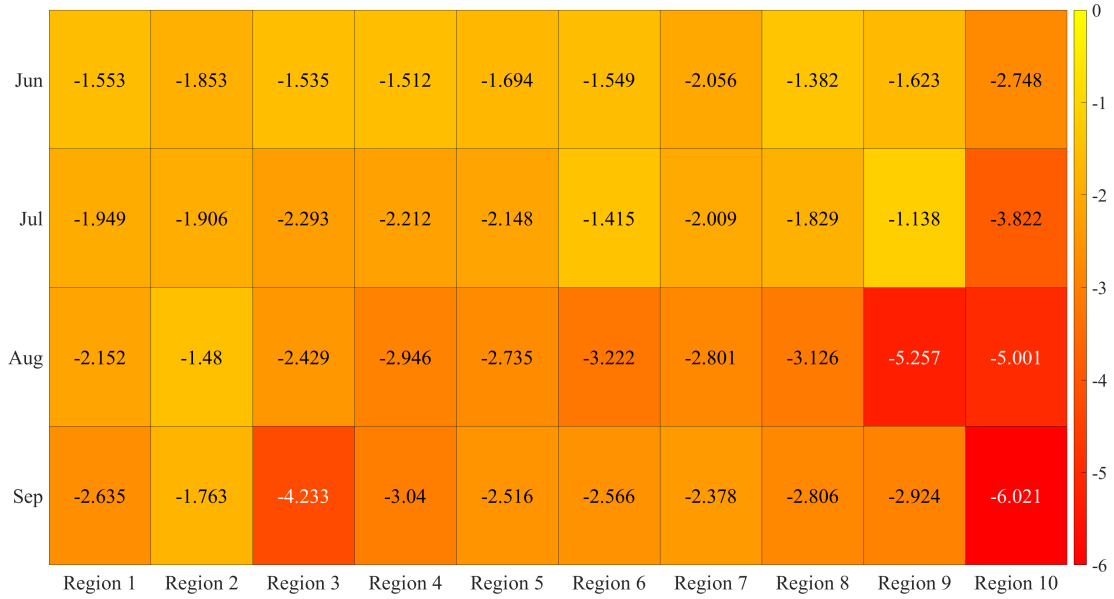

Figure S1. The changes in RMSE at every region and month. The more negative changes (darker colors) show better results since it means that the RMSE has decreased the most in that scenario.

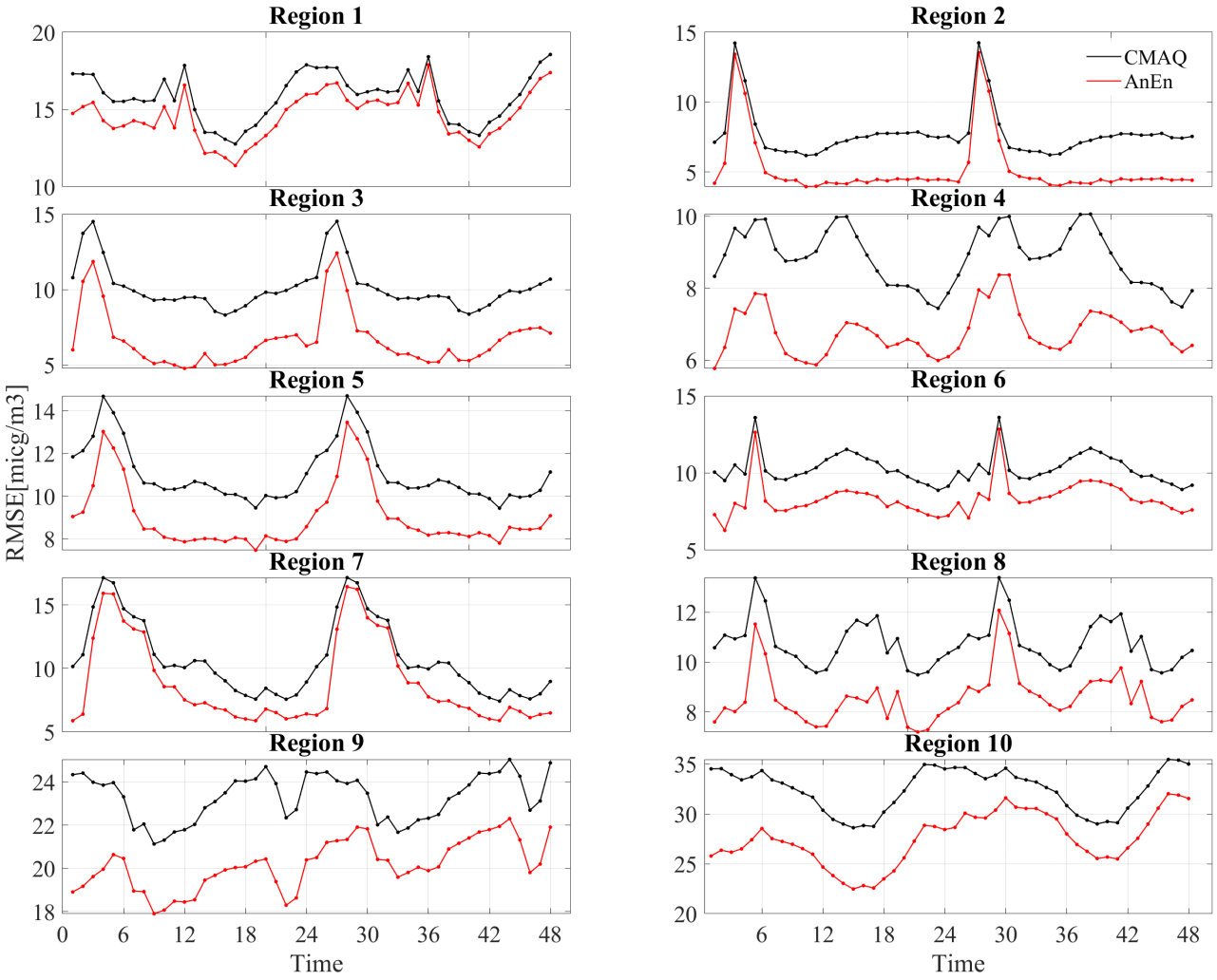

Figure S2. RMSE for CMAQ deterministic forecasts in black, and for AnEn corrected forecasts in red, calculated over all stations within each region during the study time.

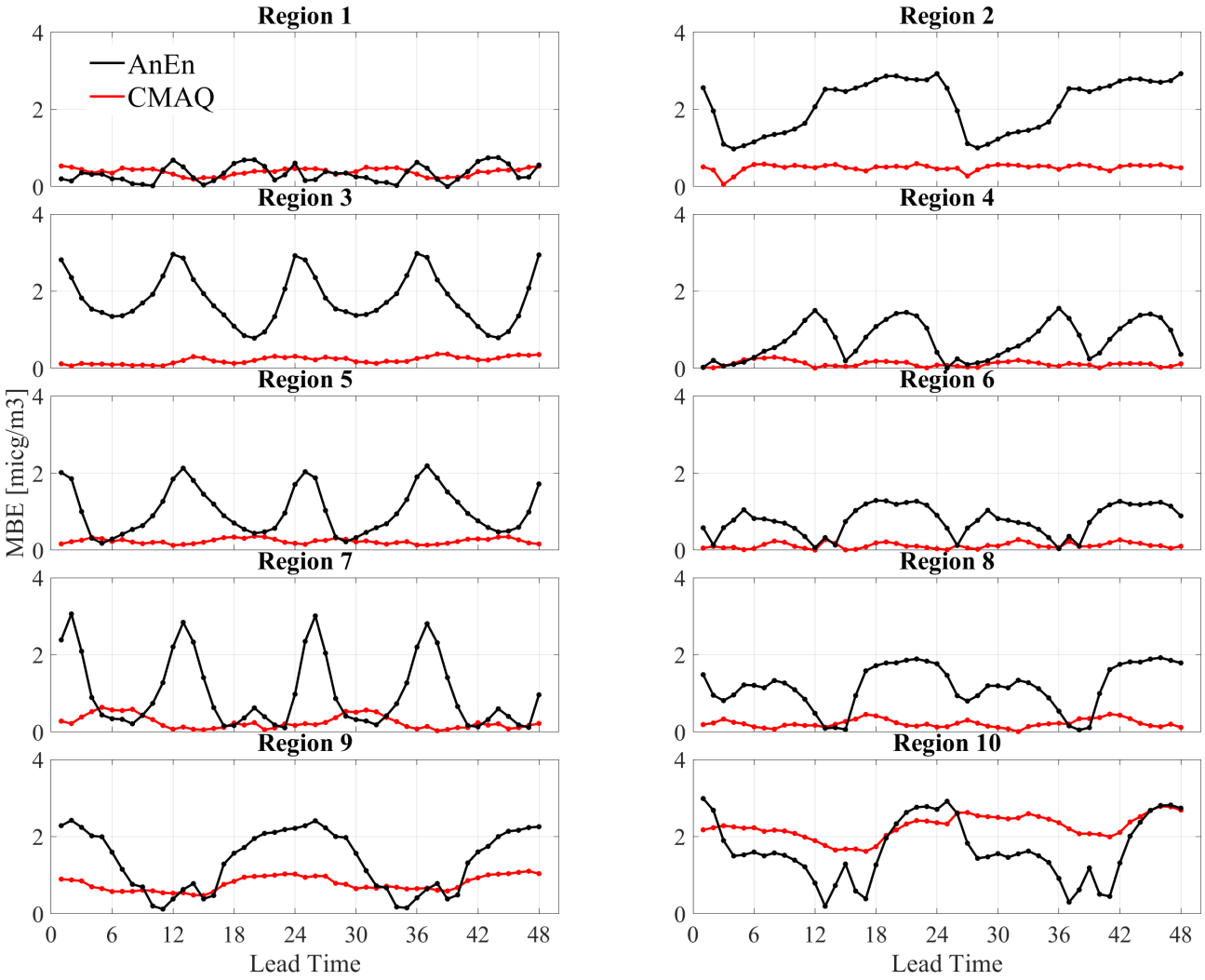

Figure S3. Same as Fig. S2, but for Mean Bias Error.

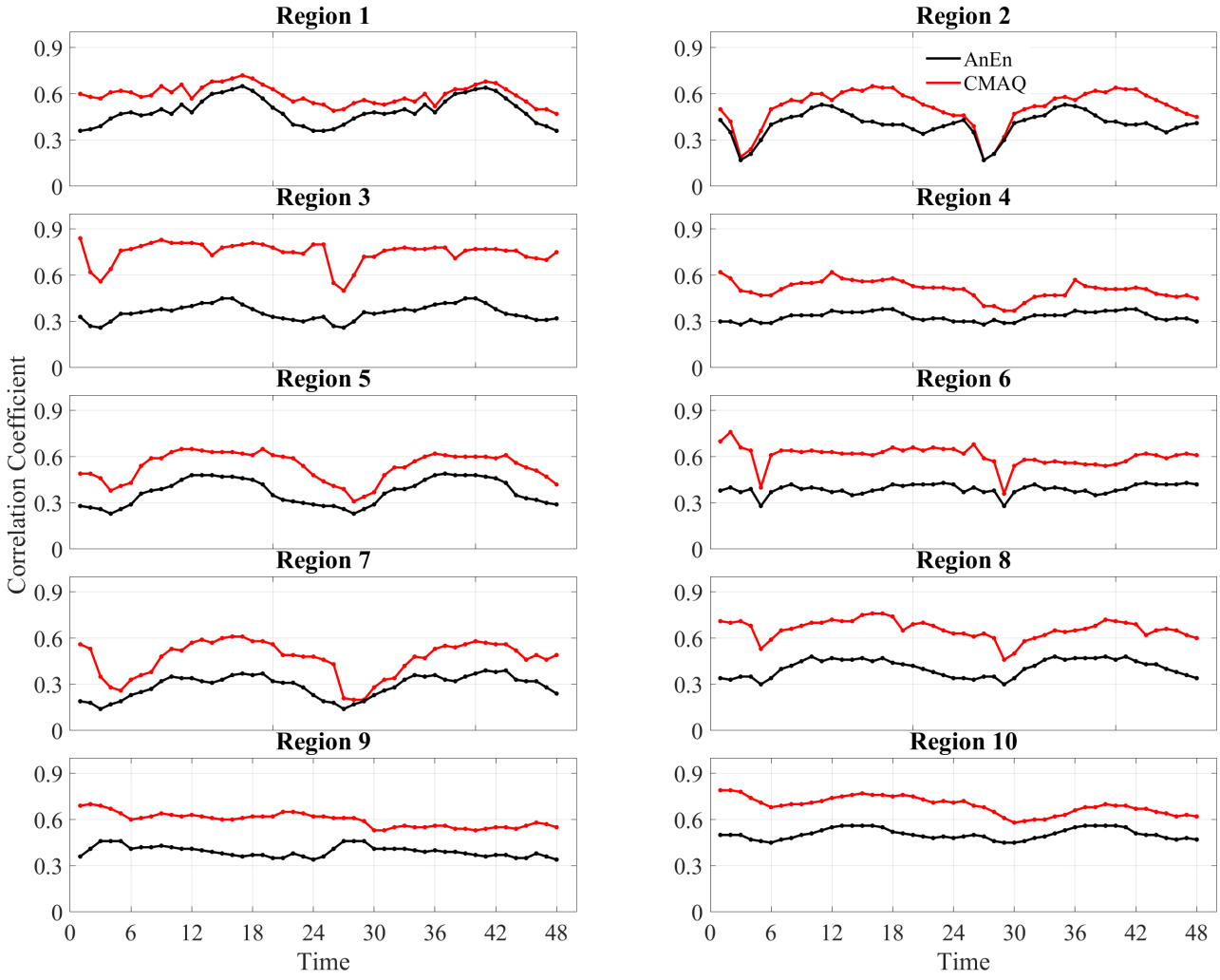

Figure S4. Same as Fig. S2, but for Correlation Coefficient.

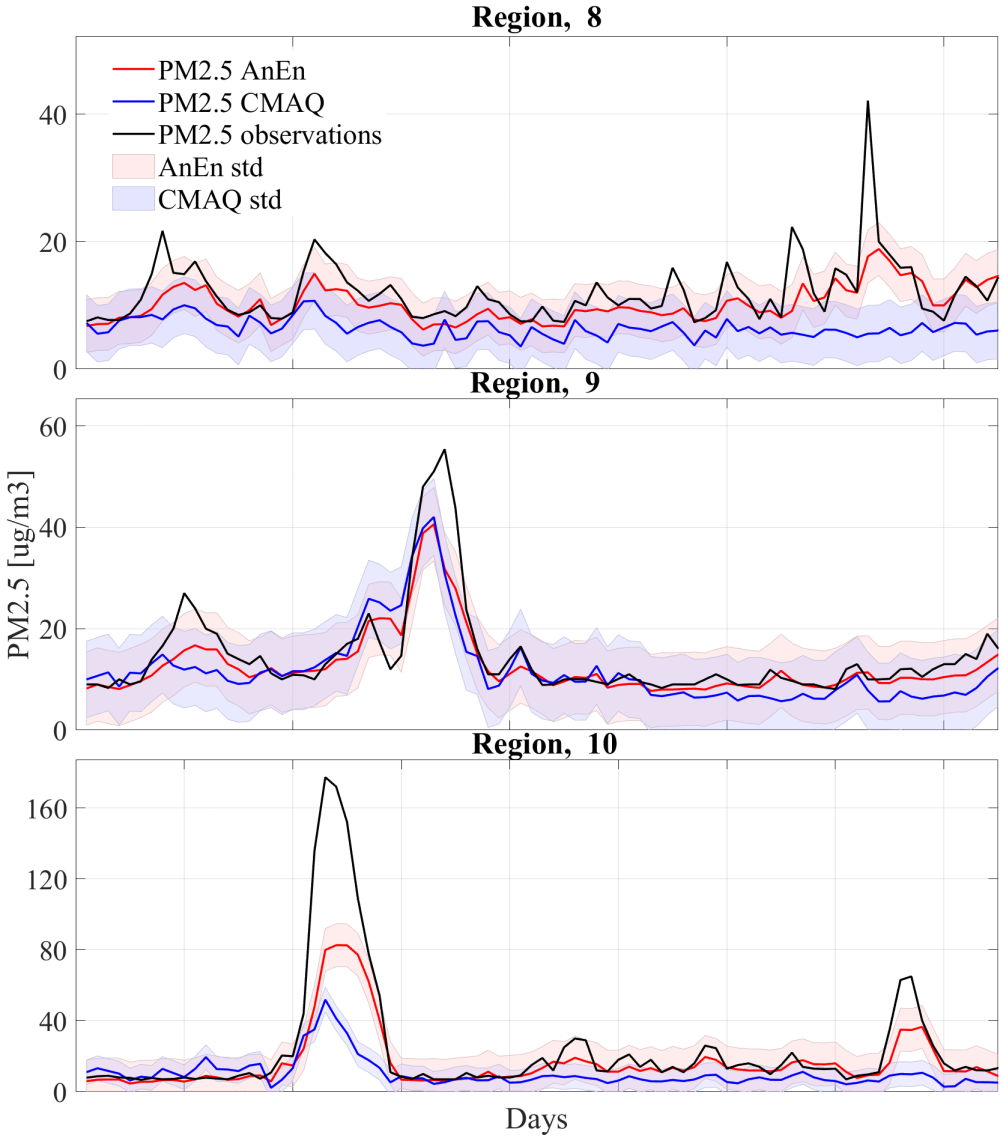

Figure S5. $PM_{2.5}$ concentrations on days when observed $PM_{2.5}$ levels exceeded the 80th percentile in regions 8, 9, and 10. $PM_{2.5}$ levels from AQS measurements, CMAQ forecasts, and forecasts after AnEn post-processing are shown in black, blue, and red lines, respectively.