Statistical Method-Based Multi-Model Ensemble Forecasting for North American Hazardous Air Quality

Yunyao Li\textsuperscript{1}, Daniel Tong\textsuperscript{1}, Peewara Makkaroon\textsuperscript{1}, Timothy DelSole\textsuperscript{2}, Youhua Tang\textsuperscript{3}, Patrick Campbell\textsuperscript{1}, Barry D. Baker\textsuperscript{3}, Mark Cohen\textsuperscript{3}, Anton Darmenov\textsuperscript{4}, Ravan Ahmadov\textsuperscript{5}, Edward J. Hyer\textsuperscript{6}, and Peng Xian\textsuperscript{7}

\textsuperscript{1}George Mason University
\textsuperscript{2}George Mason University and Center for Ocean-Land-Atmosphere Studies
\textsuperscript{3}National Oceanic and Atmospheric Administration (NOAA)
\textsuperscript{4}NASA
\textsuperscript{5}University of Colorado at Boulder and NOAA
\textsuperscript{6}Naval Research Laboratory
\textsuperscript{7}United States Naval Research Laboratory

July 23, 2023

Abstract

Wildfires can cause hazardous air quality conditions and affect human health. Accurate prediction of wildfire air quality is challenging due to uncertainties in fire emission, plume rise, etc. Ensemble forecasting techniques have been increasingly used to improve the predictability of wildfire aerosols. In this study, we aim to explore several ensemble creation methods to improve the ensemble forecast compared to the traditional ensemble mean. Our results show that the ensemble mean outperforms individual models in predicting fine particles (PM2.5) when compared to AirNow observations. To further improve the ensemble forecast, we employ various statistical methods, including multiple linear regression, ridge regression, weighted regression, and quantile regression to develop weighted ensembles. These weighted ensembles are capable of reducing model systematic error, e.g., fractional bias was 25\% to 53\% lower than the ensemble mean in the major fire regions. Furthermore, the weighted ensemble using ridge regression increased the hit rate by 17\% and reduced the false alarm rate by 72\% compared to the ensemble mean. This demonstrates that the weighting of the members comprising the ensemble is an effective method for reducing forecast uncertainty and improving the accuracy of air quality forecasting during wildfire events. The quantile and weighted regression methods improved the forecast of extreme air quality events. The PM2.5 exceedance hit rate is improved by 55\% compared to the ensemble mean. Our findings provide insights into the development of advanced ensemble methods for wildfire air quality forecast offering support for decision-making to protect public health.

Hosted file

Statistical Method-Based Multi-Model Ensemble Forecasting for North American Hazardous Air Quality

Yunyao Li 1,2,3, Daniel Tong 1,2, Peewara Makkaroon 1, Timothy DelSole 1, Youhua Tang 2, 3, Patrick Campbell 2,3, Barry Baker 3, Mark Cohen 3, Anton Darmenov 4, Ravan Ahmadov 5, Edward Hyer 6, Peng Xian 6

1Department of Atmospheric, Oceanic and Earth Sciences, George Mason University, Fairfax, VA, USA
2Center for Spatial Information Science and Systems, George Mason University, Fairfax, VA, USA
3Air Resources Laboratory, National Oceanic and Atmospheric Administration, College Park, MD 20740, USA
4National Aeronautics and Space Administration Goddard Space Flight Center, Greenbelt, USA
5Cooperative Institute for Research in Environmental Sciences, University of Colorado at Boulder, Boulder, CO, USA
6Marine Meteorology Division, Naval Research Laboratory, Monterey, CA, USA

*Corresponding authors: Yunyao Li (yli74@gmu.edu); Daniel Tong (qtong@gmu.edu)

Key Points:

- A real-time ensemble forecasting system for wildfire air quality has been developed, tested, and implemented.
- The ensemble forecast consistently works better than any of the individual models.
- Several weighted ensemble methods can further improve the ensemble forecast from the unweighted ensemble mean.
Abstract

Wildfires can cause hazardous air quality conditions and affect human health. Accurate prediction of wildfire air quality is challenging due to uncertainties in fire emission, plume rise, etc. Ensemble forecasting techniques have been increasingly used to improve the predictability of wildfire aerosols. In this study, we aim to explore several ensemble creation methods to improve the ensemble forecast compared to the traditional ensemble mean. Our results show that the ensemble mean outperforms individual models in predicting fine particles (PM$_{2.5}$) when compared to AirNow observations. To further improve the ensemble forecast, we employ various statistical methods, including multiple linear regression, ridge regression, weighted regression, and quantile regression to develop weighted ensembles. These weighted ensembles are capable of reducing model systematic error, e.g., fractional bias was 25% to 53% lower than the ensemble mean in the major fire regions. Furthermore, the weighted ensemble using ridge regression increased the hit rate by 17% and reduced the false alarm rate by 72% compared to the ensemble mean. This demonstrates that the weighting of the members comprising the ensemble is an effective method for reducing forecast uncertainty and improving the accuracy of air quality forecasting during wildfire events. The quantile and weighted regression methods improved the forecast of extreme air quality events. The PM$_{2.5}$ exceedance hit rate is improved by 55% compared to the ensemble mean. Our findings provide insights into the development of advanced ensemble methods for wildfire air quality forecast offering support for decision-making to protect public health.

Plain Language Summary

Wildfires can cause harmful air quality conditions and have adverse effects on human health. However, predicting air quality during wildfires is challenging due to uncertainties in factors like fire emissions and plume rise. Ensemble forecasting techniques are being increasingly used to enhance the accuracy of wildfire aerosol predictions. In this study, our goal was to explore different ensemble creation methods to improve upon the traditional ensemble mean approach. Our results indicate that the ensemble mean performs better than individual models. We employed various statistical methods, including multiple linear regression, ridge regression, weighted regression, and quantile regression, to develop weighted ensembles. These fractional bias of the weighted ensemble is 25% to 53% lower than the ensemble mean. Moreover, the weighted ensemble improved the hit rate by 17% and reduced the false alarm rate by 72%. This highlights the effectiveness of assigning weights to individual ensemble members in reducing forecast uncertainty and enhancing the accuracy of air quality predictions during wildfires. The quantile and weighted regression methods specifically improved the forecasting of extreme air quality events. Overall, our findings contribute valuable insights into the development of advanced ensemble methods for wildfire air quality forecasting, providing support for decision-making to protect public health.
1 Introduction

Wildfires are a significant contributor to atmospheric aerosols and trace gases, resulting in hazardous air quality and adverse health effects. Previous studies have shown that exposure to wildfire smoke is associated with all-cause mortality and respiratory morbidity (Koning et al., 1985; Reid et al., 2016; Cascio, 2018). The global average mortality attributable to landscape fire smoke exposure was estimated to be 339,000 deaths annually (Johnston et al., 2012). O’Neill et al. (2021) examined the regional health impacts of the 2017 Northern California wildfires and estimated 83 excess deaths from exposure to PM$_{2.5}$ (i.e., particles having an aerodynamic diameter less than 2.5 μm), of which 47% were attributable to wildfire smoke during the smoke episode. Liu et al. (2021) assessed the health impact of the 2020 Washington State wildfire smoke episode; they estimated 38 more all-causes mortality cases and 15 more respiratory mortality cases.

Accurate forecasting of air quality during wildfire events is crucial for public health management and emergency response, including early warnings and air quality alerts, but it remains a challenging task due to various uncertainties in the forecasting systems. These include uncertainties in fire emissions (Liu et al., 2020; Pan et al., 2020), plume rise calculations (Briggs 1969; Freitas et al., 2007; Stein et al., 2009; Rio et al., 2010; Sofiev et al., 2012; Paugam et al., 2016; Vernon et al., 2018; Zhu et al., 2018; Li et al., 2023), and other model inputs/processes (Delle Monache and Stull, 2003; Kumar et al., 2020; Li et al., 2020). O’Neill et al (2023) provides a summary of the current state and challenges of operational modelings on wildfire smoke. Ye et al. (2021) evaluated the smoke emissions and plume forecasts from 12 state-of-the-art air quality forecasting systems during the Williams Flats fire in 2019 and found large differences between the models. Specifically, they found that organic carbon emissions used by different models varied by a factor of 20 to 50.

Ensemble forecasting techniques have been increasingly used to improve the predictability of extreme air quality episodes. Gilliam et al. (2015) used the Short-Range Ensemble Forecast system to drive the four-dimensional data assimilation in the Weather Research and Forecasting (WRF)-Community Multiscale Air Quality (CMAQ) model, by perturbing initial conditions to simulate ozone chemistry and transport. Their study revealed that the differences in WRF-simulated meteorological fields can cause a significant difference in simulated ozone mixing ratios, which could reach up to 10–20 ppb or 20–30% in areas with high pollution levels. Sessions et al. (2015) and Xian et al. (2019) developed and evaluated the International Cooperative for Aerosol Prediction (ICAP) multi-model ensemble (MME), a global operational aerosol multi-model ensemble for the aerosol optical depth (AOD) forecast. The ICAP-MME reduces forecast bias and increased correlation with observation, and its performance has been stable compared to individual models over the years. Li et al. (2020) used an ensemble forecast to predict surface PM$_{2.5}$ during the 2018 California Camp Fire event using the NOAA Hybrid Single-Particle Lagrangian Integrated Trajectory (HYSPLIT) dispersion model with different emissions, plume heights, and model setups. They found that the ensemble mean outperformed any of the ensemble members, indicating that ensemble forecasting could reduce forecast uncertainty. Makkaroon et al. (2023) successfully developed a multi-model ensemble forecast system that effectively simulated the 2020 western US "Gigafire", with the ensemble mean outperforming individual models. These studies highlights the potential of ensemble forecasting to improve the predictability of air quality during wildfires.
While multi-model ensemble often outperforms single model some challenges remain. Ensemble forecasting does not work best all the time. For example, the models used to create the ensemble may not be diverse enough. If the models used are too similar, the ensemble may not capture the full range of uncertainties and variability associated with different model inputs and assumptions. In addition, if the individual models used to create the ensemble are biased, the ensemble may also exhibit systematic bias.

In this study, we aim to develop a real-time regional multi-model ensemble forecast over the Contiguous United States (CONUS) that can enhance the predictability of extreme air quality events, such as wildfires. To achieve this, we propose several statistical-based methods to optimize the ensemble design and compare the performance of these new methods to that of the simple ensemble mean method used in previous studies (e.g., Li et al., 2020; Makkaroon et al., 2023). We evaluated the performance of these ensemble forecasts for the entire year of 2022. The ensemble design methods and the evaluation methods are described in Section 2, and the results are discussed in Section 3.
2 Materials and Methods

2.1 Wildfires in 2022

The study period of this paper is the year of 2022. According to the National Interagency Fire Center, wildfires in 2022 burned a total of 3,066,377 hectares across the United States. Figure 1 displays the annual and monthly total fire radiative energy (FRE) from Global Biomass Burning Emissions Product (GBBEPx; Zhang et al., 2012, 2014, 2019), which is highly correlated with fire emissions, across the 10 U.S. Environmental Protection Agency (EPA) regions for 2022. In the eastern U.S., biomass burning emissions were mainly concentrated in the southeastern states (Region 4), including Florida, Georgia, and Alabama. Although the Southeast fires, mostly agricultural fires and prescribed burning, affected a large area, the total fire radiative energy was not as high as that of the western wildfires. The peak fire period for Region 4 occurred in early spring, especially in March when 6,281 TJ of fire energy was released in one month. In the central U.S., fire emissions were primarily located in Arkansas, Kansas, Texas, and New Mexico (Region 6 and 7). Texas experienced its most significant wildfire year within the last decade, with over 12,000 wildfires burning more than 650,000 acres across the state. In New Mexico, several large fires started in April and May, such as the Hermits Peak and Calf Canyon Fires, which burned 341,735 acres. The peak fire time for the central U.S. is in spring, especially in April and May, when a total of 41,634 TJ of fire energy was released in these regions within two months. In the western U.S., fires were primarily located in Regions 9 and 10, with the peak fire period occurring in the summer, especially in September, when 22,804 TJ of fire energy was released in one month. Overall, the strongest fire energy was recorded in September, with 33,631 TJ, followed by May, with 30,330 TJ.
Figure 1. The annual and monthly total fire radiative energy, which is highly correlated with fire emissions, across the 10 U.S. EPA regions for 2022.

2.2 Description of Ensemble Members

The air quality forecast ensemble in this study was developed using both regional and global chemical transport models, which included the National Oceanic and Atmospheric Administration (NOAA) High-Resolution Rapid Refresh-Smoke (HRRR-Smoke), Global Ensemble Forecast System Aerosols (GEFS-Aerosols), NOAA National Air Quality Forecasting Capability (NAQFC), NASA Goddard Earth Observing System (GEOS), and Navy Aerosol Analysis and Prediction System (NAAPS). These models range from simple smoke tracer models to full air quality models with gas/aerosol chemistry, from high-resolution regional to coarse reso. Generating an ensemble based on these widely different forecast models helps to take advantage of the strengths of the models that they can offer. The forecast models here used to create the ensemble cover a broad range of emission datasets, as well as different plume rise schemes. The study utilizes the 12-36 hour surface PM$_{2.5}$ forecasts initialized at 12 UTC (forecast hour: 00-23 UTC the next day) for all five models. Each model is briefly described below.
2.2.1 HRRR-Smoke

HRRR-Smoke (Ahmadov, et al., 2017; Dowell et al., 2022) is a three-dimensional coupled weather-smoke model that enables the simulation of mesoscale flows and smoke dispersion over complex terrain, in the boundary layer, and aloft at the 3 km spatial resolution over the Continental United States (CONUS) domain in real-time. The HRRR-Smoke is based on the Weather Research and Forecasting (WRF) model, which uses the Thompson aerosol microphysics scheme (Thompson and Eidhammer, 2014) and the Mellor-Yamada-Nakanishi-Niino (MYNN) planetary boundary layer (PBL) scheme (Olson et al., 2019). The 13.5 km Rapid Refresh coupled with Smoke (RAP-Smoke) model provides the boundary conditions of meteorology and smoke to HRRR-Smoke. The HRRR Data-Assimilation System provides initial conditions (ensemble mean) and a background ensemble for meteorological data assimilation. The model also assimilates radar data every 15 minutes.

HRRR-Smoke ingests the satellite fire radiative power data (FRP) from the Suomi-NPP, NOAA-20, and MODIS Terra/Aqua satellites in real time. The FRP data are used to estimate the smoke (primary particulate matter) emissions from wildland fires. It should be noted that HRRR-Smoke does not include any non-fire emissions (e.g. anthropogenic emissions) and gas/aerosol chemistry. Thus, the model is designed to forecast PM$_{2.5}$ in conditions, when smoke is a dominant source of air pollution in a given region. Currently, HRRR-Smoke is an operational forecast model, maintained by NOAA National Centers for Environmental Prediction (NCEP).

2.2.2 GEFS-Aerosol

GEFS-Aerosols (Hamill et al., 2011) is a global atmospheric composition model developed by the National Center for Environmental Prediction (NCEP) in collaboration with the NOAA Global Systems Laboratory (GSL), NOAA Chemical Sciences Laboratory (CSL), and NOAA Air Resources Laboratory (ARL). It is an atmospheric composition model that integrates weather and air quality. The meteorology component uses the Finite Volume Cubed Sphere (FV3)-based Global Forecast System (GFS) version 15, while the atmospheric chemistry component is based on the Weather Research and Forecasting-Chemistry (WRF-Chem) model. The aerosol modules are based on the NASA Goddard Chemistry Aerosol Radiation and Transport model (GOCART). The operational GEFS-Aerosols model currently incorporates biomass burning emission data from GBBEPx and global anthropogenic emission data from the Community Emission Data System (CEDS, Jeong et al., 2022). Wildfire smoke plumes are calculated using a one-dimension (1-D) time-dependent cloud module from the HRRR-Smoke model (Freitas et al., 2007). This study utilized the GEFS-Aerosols global PM$_{2.5}$ forecasts at a horizontal resolution of 0.25° × 0.25°.

2.2.3 NAQFC

NOAA's operational National Air Quality Forecasting Capability (NAQFC), which has a horizontal spatial resolution of 12 km and 35 vertical layers, uses CMAQ version 5.3.1 driven by NOAA's latest operational FV3-GFSv16 meteorology via the NOAA-EPA Atmosphere–Chemistry Coupler (NACC, Campbell et al., 2022), a meteorological preprocessor adapted from the existing Meteorology–Chemistry Interface Processor (MCIP). The chemical gaseous boundary conditions are based on static, global GEOS-Chem simulations, while aerosol boundary conditions for smoke and dust are dynamically updated from NOAA’s operational GEFS-Aerosols model (Section 2.2.2). NAQFC employs GBBEPx for biomass burning
emissions, NEIC 2016v1 for anthropogenic emissions, and the Biogenic Emission Inventory System version 3.6.1 (BEISv3.6.1; Vukovich and Pierce, 2002; Schwede, 2005) with the Biogenic Emission Landuse Dataset version 5 (BELD5) for inline biogenic volatile organic carbon (BVOC) emissions. The model uses the Briggs (1969) plume rise algorithm to compute wildfire smoke plumes. See Campbell et al. (2022) for more information on the NAQFC system components and related references.

2.2.4 GEOS

The GEOS (Gelaro et al., 2017, Buchard et al., 2017; Randles et al., 2017) system is developed by NASA’s Global Modeling and Assimilation Office (GMAO). This study used the GEOS Forward Processing system (GEOS-FP, version 5.27.1). The GEOS-FP generates analyses, assimilation products, and ten-day forecasts in near-real time. GEOS-FP is built around the GEOS Atmospheric General Circulation Model (AGCM), the GEOS atmospheric data assimilation system (hybrid–4DEnVar ADAS), and aerosol assimilation (Randles et al., 2017). Aerosols are an integral component of the model physics (Buchard et al., 2017) and are simulated with the Goddard Chemistry, Aerosol, Radiation, and Transport model (GOCART; Chin et al., 2002; Colarco et al., 2010). Fire emissions come from the Quick Fire Emissions Dataset (QFED; Darmenov and da Silva, 2015) and leverage low-latency MODIS fire locations and fire radiative power (FRP, Collection 6) data. Emissions from fires are distributed in the Planetary Boundary Layer (PBL). Anthropogenic emissions are from the Emissions Database for Global Atmospheric Research (EDGAR) and Hemispheric Transport of Air Pollution (HTAP) inventories. BVOC emissions are from the Model of Emissions of Gases and Aerosols from Nature (MEGAN).

2.2.5 NAAPS

NAAPS (Lynch et al., 2016) is developed at the Marine Meteorology Division of the Naval Research Laboratory (NRL) and provides an operational forecast of 3D atmospheric anthropogenic fine and biogenic fine aerosols, biomass burning smoke, dust, and sea salt concentrations on a spatial resolution of $0.333^\circ \times 0.333^\circ$. The current NAAPS is driven by global meteorological fields from the Navy Global Environmental Model (NAVGEM), an operational global weather prediction system developed by the United States Navy (Hogan et al., 2014). NAAPS uses a biomass burning source from the Fire Locating and Modeling of Burning Emissions (FLAMBE) inventory, which is based on near-real time MODIS fire hotspot data (Reid et al., 2009). The wildfire smoke at emission is distributed uniformly through the bottom 4 layers within the PBL. The NAAPS analysis is constrained by the assimilation of MODIS AOD (Zhang et al., 2008; Hyer et al., 2011).

2.3 Description of Observations

The hourly ground PM$_{2.5}$ observations from the U.S. EPA AirNow network for the year of 2022 are used to evaluate the surface air pollution predictions in this study. The real-time AirNow measurements are collected by the state, local, or tribal environmental agencies using federal references or equivalent monitoring methods approved by EPA. It contains air quality data for more than 500 cities across the U.S., as well as for Canada and Mexico.
2.4 Evaluation method

2.4.1 Random walk

We use the random walk method proposed by Delsole and Tippett (2016) to assess the performance of the ensemble forecast and each individual model. Specifically, suppose forecasts A and B are compared N times. Whenever A is more skillful than B, a step in the positive direction is taken; otherwise, a step in the negative direction is taken. Let K represents the number of times that forecast A outperforms forecast B. The net distance \( d_N \) traveled by the random walk is:

\[
d_N = K - (N - K) = 2K - N
\]  

Fractional bias (FB) is used for deciding the most skillful forecast of a single event (i):

\[
FB_i = 2 \times \frac{|O_i - M_i|}{O_i + M_i}
\]  

where O is the AirNow observation, and M is the model forecast. A significance test \( K_\alpha \) is conducted to show if A is significantly better \((K > K_\alpha)\) or worse \((K < N - K_\alpha)\) than B. \( K_\alpha \) can be approximated as:

\[
K_\alpha = \left[ \frac{N}{2} - z_\alpha \sqrt{\frac{N}{4} - \frac{1}{2}} \right]
\]  

where \( z_\alpha \) is the value for which a standardized Gaussian is exceeded with probability \( \alpha = 5\% \), and \([x]\) denotes ceiling function that maps x to the smallest integer greater of equal to x.

2.4.2 Categorical Metrics

The statistical metrics like fractional bias mentioned above have limitations in evaluating the model performance of extreme events, such as wildfires. To address this, categorical metrics can be used to measure the model’s ability to predict exceedance events. In this study, we used three categorical metrics: the area hit rate (aH), the area false alarm rate (aFAR), and the weighted success index (WSI), as described by Kang et al. (2007). These metrics were based on the 24-hour PM\(_{2.5}\) exceedance level defined by the US EPA National Ambient Air Quality Standards (NAAQS), which is 35 \( \mu g/m^3 \) (U.S. EPA, 2020).

The area false alarm rate (aFAR) and area hit rate (aH) were calculated based on paired observed (O) and predicted (M) PM\(_{2.5}\) exceedances by considering three possible scenarios: a forecasted exceedance that is not observed (a); a forecasted exceedance that is observed (b); and an observed exceedance that is not forecasted (c). The aH and aFAR values are determined by matching observed and forecasted exceedances within a designated area surrounding the observation locations. In the present study, we used an area of 0.5° × 0.5° centered at each AirNow site’s location.

\[
aFAR = \left( \frac{Aa}{Aa + Ab} \right) \times 100\%
\]

\[
aH = \left( \frac{Ab}{Ab + Ac} \right) \times 100\%
\]

where \( Aa \) is the number of forecast area exceedances that were not observed (false alarms); \( Ab \) is the number of cases where an observed exceedance corresponds to a forecast exceedance within the designated area of 0.5° × 0.5° centered at the monitor location; \( Ac \) is the number of observed
exceedances that are not forecast within the designated area centered at the monitor location. The\n\(aFAR\) (4) refers to the percentage of false alarms if a forecasted exceedance is not observed\nwithin the designated area. The area hit rate \(aH\) (5) refers to the percentage of hits if a forecasted\nexceedance is observed within the designated area. The \(aFAR\) and \(aH\) both range from 0-100%.\nIf a model performs well, the misses (\(Ac\)) will be low, and the hits (\(Ab\)) will be high, resulting in\nhigh \(aH\). In contrast, if a model performs poorly, the false positives (\(Aa\)) will be high and the hits\n(\(Ab\)) will be low, resulting in high \(aFAR\).

The weighted success index (WSI) gives credit for observation (O) or prediction (M) that\nare close to the threshold (T).

\[
WSI = \frac{b + \sum_i^a IP}{a + b + c} \times 100%
\]

(6)

\[
IP = \begin{cases} 
M - fO & \text{if } O < T < M < fO \\
M - fT & \text{if } O < M < T < fO \\
O - fM & \text{if } M < T < O < fM 
\end{cases}
\]

(7)

The choice of \(f\) is empirical and is based on rules of thumb (Hanna 2006). Analysis of PM\(_{2.5}\)\nresults for 2022 has shown that about 80% of the difference between observation and prediction\nis within a factor of 2; thus, in this study, \(f\) is set to 2.

### 2.5 Ensemble design

In this section, we will describe several methods we used to create the ensemble using the\nfive models introduced in section 2.2. In the following discussion, we will refer to these models\nas model-1 through 5 instead of explicitly mentioning their names. This approach allows us to\nfocus on evaluating the performance of the ensemble forecasts, which is the primary aim of this\nstudy. It is worth noting that we intentionally rearranged the order of the models cited in section\n2.2 differently from their sequence in the results section.

#### 2.5.1 Ensemble Mean

The multimodel average (MMA) is a widely used technique for combining the forecasts\nof multiple models into a single consolidated prediction. This method involves taking the\naverage of the results obtained from each model (\(M_j\)) with equal weights. The multimodel mean\nis particularly useful when there is no a priori reason to favor one model over another, and all\nmodels have comparable skill levels.

\[
\bar{M}_{MMA} = \sum_{j=1}^{S} \frac{1}{S} M_j
\]

(8)

where \(S\) represents the total number of models which is 5 in this study.
2.5.2 Weighted Ensemble

The weighted ensemble is a method for combining multiple models by assigning different weights to each model. This approach can be more effective than the simple multimodel mean, as it allows for a more sophisticated and nuanced consolidation of the model outputs. In this section, we will introduce several statistical methods to calculate the model weights ($\beta$).

$$\bar{M} = \sum_{j=1}^{S} \beta_j M_j + \beta_0$$  \hspace{1cm} (9)

To determine the weight for each model, we partitioned the data for the year 2022 into training and testing sets. To ensure the independence of the training and testing data, we opted not to select the training data randomly. Since wildfires usually last for several days, randomly selecting data cannot guarantee the independence of the training and testing data. Instead, we used the first 9 months of data as the training set and the final 3 months as the testing set. Due to computational limitations (space and time), we were only able to analyze one year of data, which may lead to variability in the calculated weights for each model. The purpose of this paper is to introduce and test various weighted ensemble approaches for air quality forecasting. Longer training and testing periods are required before implementing a weighted ensemble in operational forecasting, in order to thoroughly investigate its performance and determine the optimal weights for each model.

2.5.2.1 Multi-linear Regression (MLR)

Multiple linear regression is a statistical technique that can be employed to calculate the weights for each model in the ensemble to improve the accuracy of the ensemble forecast by minimizing the error between the observation ($O$) and the weighted multimodel prediction:

$$\hat{\beta}_{MRL} = \arg \min_{\beta} \left( \sum_{i=1}^{N} \left( O_i - \beta_0 - \sum_{j=1}^{S} \beta_j M_{ij} \right)^2 \right)$$  \hspace{1cm} (10)

where $N$ is the total number of observations.

However, the multiple linear regression has an overfitting issue. Overfitting is a common problem in statistical modeling, machine learning, and data analysis. It occurs when a model is trained too well on the training data, to the point where it starts to fit the noise in the data instead of the underlying pattern or relationship. As a result, the model performs poorly on new, unseen data because it has become too specialized for the training data. Overfitting can be addressed by using regularization techniques, such as ridge regression (section 2.5.2.2), that constrain the size of the model parameters or coefficients.

2.5.2.2 Ridge Regression (RR)

The ridge regression (Hoerl and Kennard, 1970) is a technique used to reduce overfitting issues in linear regression models. It does so by adding a penalty term to cost function that constrains the size of the weights. The penalty term is proportional to the square of the weights, so the larger the weights, the larger the penalty:

$$\hat{\beta}_{RR} = \arg \min_{\beta} \left( \sum_{i=1}^{N} \left( O_i - \beta_0 - \sum_{j=1}^{S} \beta_j M_{ij} \right)^2 \right) + \lambda \sum_{j=1}^{S} \beta_j^2$$  \hspace{1cm} (11)
where λ is the ridge parameter. We tested the λ value from 0 to 1000 to find the best λ for ridge regression. The first 20 days in each month are used to train the data using Eq (11), and the last 10 days are used to find the best λ. This method shrinks the weights towards zero and balances the contribution of the individual models thus improving the generalization performance of the ensemble. By constraining the weights, ridge regression can produce a more robust and stable model, especially when the number of predictors is large and the data is noisy. It has been found to be a useful method in climate ensemble studies (DelSole et al., 2007; Pena and van den Dool, 2008).

2.5.2.3 Quantile Regression (QR)

The traditional regression method estimates the conditional mean of the response variable across values of the predictor variables. However, this method tends to favor the mean state, which is good for forecasting general events but not suitable for extreme events. In the case of hazardous air quality forecasts, it can lead to underestimations of pollution levels. To address this issue, we use quantile regression (Koenker, 2005), an approach similar to traditional linear regression but with quantile-dependent regression coefficients:

\[ M_{QR} = \sum_{j=1}^{S} \beta_{1,q} M_j + \beta_{0,q} \] (12)

where \( q \) represents the quantile ranging from 0 to 1. In this paper, we use \( q=0.9 \) to give more credit to the top 10% of events (use the 90th percentile of data to determine the beta coefficients). The quantile regression coefficients are estimated by minimizing the sum of asymmetrically weighted absolute deviations:

\[
\hat{\beta}_{QR} = \arg \min_{\beta} \left( \sum_{j:M_{iq} \geq M_{q}} q \left| O_i - \beta_{0,q} - \sum_{j=1}^{S} \beta_{1,q} M_{ij} \right| + \sum_{j:M_{iq} < M_{q}} (1 - q) \left| O_i - \beta_{0,q} - \sum_{j=1}^{S} \beta_{1,q} M_{ij} \right| \right) \] (13)

Quantile regression is particularly useful for analyzing extreme events (Herrera et al., 2018; Khan et al., 2019).

2.5.2.4 Weighted regression (WR)

Weighted regression is another statistical method that can be used to address the issue of extreme events. Unlike traditional regression methods, weighted regression assigns different weights to data points. The weights are used to give more importance to certain data points that are more important to the analysis:

\[
\hat{\beta}_{WR} = \arg \min_{\beta} \left( \sum_{i=1}^{N} W_i \left( O_i - \beta_0 - \sum_{j=1}^{S} \beta_j M_{ij} \right)^2 \right) \] (14)
Here, to increase the impact of extreme events in the regression analysis, we assign a weight of 10 to cases with daily PM$_{2.5}$ concentration higher than 20 $\mu$g/m$^3$ (80% of the total observations), and a weight of 1 to other points, which gives more importance to polluted days:

$$W_i = \begin{cases} 
10, & \text{if } O_i > 20\mu g/m^3 \\
1, & \text{otherwise}
\end{cases}$$(15)
3 Results and Discussion

In this section, we first evaluate the performance of the ensemble mean and compare it to the performance of each individual model. Subsequently, we compare the results of the unweighted ensemble mean with the weighted ensemble approaches described in Section 2.5.2. It should be noted that most of these models are not designed to simulate PM$_{2.5}$ for all the AirNow sites in the U.S. due to incomplete representation of the air pollutants and chemistry.

3.1 Comparison of the multi-model mean with individual models

Figure 2 presents the annual mean surface PM$_{2.5}$ concentration (contour) predicted by models 1 to 5 and the ensemble mean (multimodel average, MMA) in comparison with the AirNow observations (dots). The results from different models varied substantially, highlighting the high degree of uncertainty associated with wildfire air quality forecasting. Moreover, some of the models don’t simulate anthropogenic and other non-fire PM$_{2.5}$ sources (e.g. HRRR-Smoke), or deploy simplified aerosol parametrization (e.g. GEFS). Specifically, models 1, 2, and 4 overestimate the PM$_{2.5}$ in the southeast and northwest regions, while models 3 and 5 underestimate it. The ensemble mean offers a balanced view, accounting for both overestimations and underestimations, and is closer to the observations.
Figure 2. Annual mean surface PM$_{2.5}$ concentration (contour) predicted by models 1 to 5 and the ensemble mean (multimodel average, MMA) in comparison with the AirNow observation (colored circles) of the year 2022.

We compared the MMA with each individual model using the random walk method and the results are displayed in Figure 3 for each EPA region, where positive values and tendencies indicate that the individual model performs better than the ensemble mean (MMA), and negative values and tendencies indicate that the MMA is superior to the individual model. All random walk scores begin at zero. The random walk scores in regions 1, 3, 4, 10 consistently trend
downward, implying that MMA consistently outperforms individual models in this region. In regions 5, 6, 7, and 9, the scores are mostly negative with some transient positive scores in the early of the year as well as some positive tendency at the end of the year, indicating that MMA performs better than each model most of the time. For Region 2, MMA performs better than Models 1 to 4 but worse than Model-5, particularly in the first half of the year. For Region 8, the MMA is better than Model-1 and 3, worse than Model-2 all the time, worse than Model-4 in the first half of the year, better in the second half, and worse than Model-5 in the Jan-Mar and Oct-Dec. Overall, the MMA performs better than individual models, indicating that ensemble forecasts can effectively reduce forecast uncertainty.
Figure 3. Comparing MMA to individual models using the random walk method for EPA regions.

To evaluate the ability of each individual model and the ensemble mean (multimodel average, MMA) to forecast extreme events, the area hit rate, area false alarm rate, and weighted success index are calculated and shown in Table 1. Model 4 achieves the highest hit rate (48.12%), indicating that it captures the most extreme events. However, its false alarm rate is
also the highest, with approximately 93.1% of the exceedance events forecasted by Model 4 being false alarms. On the other hand, Model 3 has the lowest false alarm rate of 29.77%, but also the lowest hit rate (14.68%). The MMA has the highest weighted success index (WSI), the second-lowest false alarm rate, and the third-highest hit rate. Overall, the ensemble mean works better than the individual models in the aspect of extreme events forecast, consistent with the results of earlier studies (Li et al., 2020; Makkaroon et al., 2023).

Table 1. The area hit rate (aH), area false alarm rate (aFAR), and weighted success index (WSI) for Models 1 to 5 and the MMA for the whole year of 2022. The best results are highlighted in bold font, while the worst results are underlined.

<table>
<thead>
<tr>
<th></th>
<th>Model-1</th>
<th>Model-2</th>
<th>Model-3</th>
<th>Model-4</th>
<th>Model-5</th>
<th>MMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>aH</td>
<td>26.98</td>
<td>41.12</td>
<td>14.68</td>
<td><strong>48.12</strong></td>
<td>12.42</td>
<td>37.44</td>
</tr>
<tr>
<td>aFAR</td>
<td>83.29</td>
<td>79.70</td>
<td><strong>29.77</strong></td>
<td>93.10</td>
<td>84.17</td>
<td>77.09</td>
</tr>
<tr>
<td>WSI</td>
<td>13.06</td>
<td>20.10</td>
<td>16.47</td>
<td>6.98</td>
<td>13.82</td>
<td><strong>20.68</strong></td>
</tr>
</tbody>
</table>

Figure 4 shows the histogram of the model bias (model minus observation) for each individual model and the MMA. All models have systematic errors (consistent positive/negative bias) in different regions. Model 3 tends to underestimate PM$_{2.5}$ concentration in all regions, while Model 4 tends to overestimate it. Models 1 and 2 tend to overestimate pollution in major fire regions (R4, 6, 7, 9, and 10), and Model 5 tends to underestimate it in those regions. The MMA balances the systematic errors of the five individual models in most regions, but it still tends to overestimate in R4, 6, and 10, and underestimate in R5, 8, and 9. These results demonstrate that the MMA can improve air quality forecasting, but there is still room for improvement.
Figure 4. Histograms of the model error for each individual model and the ensemble mean.
3.2 Weighted ensemble

In this section, we evaluate the performances of four weighted ensembles as described in Section 2.5.2 and compare their performances to that of the MMA as well as each individual model.

Figure 5 displays scatter plots between AirNow surface PM$_{2.5}$ observations and different ensemble forecasts, including the MMA (in green), multi-linear regression (MLR, in magenta), and ridge regression (RR, in black) for the testing period from October to December 2022. Compared to the unweighted ensemble mean (MMA), the weighted ensemble MLR has smaller systematic errors, particularly for the major fire regions (R4, 6, 7, 9, and 10), and is closer to the observations. Table 2 shows the fractional bias for the same period. In the major fire regions, the MLR reduced the forecast bias by 25% to 53% compared to the MMA, with the largest improvements seen in Region 7 and the least in Region 10. Table 3 shows the area hit rate, false alarm rate, and weighted success index for the same testing period. The MLR increases the hit rate by 17%, significantly reduces the false alarm rate by 72%, and increases the WSI by 5% compared to the MMA. These results demonstrate that the weighted ensemble outperforms the unweighted ensemble.

The results of RR are generally comparable to those of MLR. In regions 1-7, RR has a slightly higher fractional bias than MLR, which suggests that overfitting may not be a significant issue in these regions. However, in the western coast Regions 9 and 10, RR has a lower bias than MLR, suggesting that RR might help address the overfitting problem in these large wildfire regions. Compared to MLR, RR has a slightly lower hit rate, lower false alarm rate, and lower weighted success index.

According to Figure 5, MMA, MLR, and RR all tend to underestimate the surface PM$_{2.5}$ exceedance events, particularly in the large fire regions 8, 9, and 10. Therefore, we applied quantile regression (QR) to address this issue to generate more extreme cases. The results are presented in Figure 6, Tables 2 and 3. As depicted in Figure 6, QR enables the ensemble model to predict more high PM$_{2.5}$ events compared to MLR and MMA, but sometimes it overestimates the pollution level when the actual pollution level is not high. The fractional bias of QR is higher than that of MLR and RR but still lower than that of MMA except in Region 1. As QR predicts more extreme cases, it has a much higher hit rate, which is about 55% higher than the MMA and 33% higher than the MLR. However, its false alarm rate is also higher, reaching 32.89%. QR has the highest WSI among all models, including individual models and ensemble forecasts.

Another approach to improve the ensemble forecast ability to predict extreme cases is to use weighted regression (WR) to calculate the weight for each model. In this study, a weight of 10 is assigned to the cases with PM$_{2.5}$ concentrations higher than 20 μg/m$^3$, giving more importance to polluted days. The results for WR are presented in Figure 7 and Tables 2 and 3. Figure 7 shows that for high PM$_{2.5}$ cases, the ensemble forecast using the WR method is closer to the observations than MMA and MLR, while generating fewer false high values compared to QR. The model bias of WR is about 10% higher than MLR but much lower than QR and MMA. The area hit rate of WR is higher than that of MMA, MLR, and RR but lower than that of QR. The false alarm rate of WR is slightly higher than that of MLR and RR but much lower than that of MMA and QR. Its WSI is the second-highest among all models.
Figure 5. Scatter plots between observed and forecast PM$_{2.5}$ for MMA (green), MLR (magenta), and RR (black). The solid black line represents the 1:1 ratio line for the observations and forecasts, while the dashed black lines represent the 1:2 and 2:1 ratio lines.
Figure 6. Same as Figure 5 but for QR.
Figure 7. Same as Figure 5 but for WR.

Table 2. Fractional bias for the different ensembles for the October to December 2022 testing period. (bold represents the best results, and underline represents the worst results)

<table>
<thead>
<tr>
<th></th>
<th>MMA</th>
<th>MLR</th>
<th>RR</th>
<th>QR</th>
<th>WR</th>
</tr>
</thead>
<tbody>
<tr>
<td>R01</td>
<td>0.326</td>
<td>0.257</td>
<td>0.265</td>
<td>0.353</td>
<td>0.268</td>
</tr>
<tr>
<td>R02</td>
<td>0.340</td>
<td>0.232</td>
<td>0.236</td>
<td>0.313</td>
<td>0.241</td>
</tr>
<tr>
<td>R03</td>
<td>0.353</td>
<td>0.258</td>
<td>0.262</td>
<td>0.351</td>
<td>0.285</td>
</tr>
<tr>
<td>R04</td>
<td>0.431</td>
<td>0.240</td>
<td>0.242</td>
<td>0.331</td>
<td>0.273</td>
</tr>
<tr>
<td>R05</td>
<td>0.443</td>
<td>0.250</td>
<td>0.253</td>
<td>0.307</td>
<td>0.264</td>
</tr>
<tr>
<td>R06</td>
<td>0.527</td>
<td>0.285</td>
<td>0.292</td>
<td>0.386</td>
<td>0.395</td>
</tr>
<tr>
<td>R07</td>
<td>0.567</td>
<td>0.268</td>
<td>0.271</td>
<td>0.368</td>
<td>0.299</td>
</tr>
<tr>
<td>R08</td>
<td>0.732</td>
<td>0.367</td>
<td>0.370</td>
<td>0.473</td>
<td>0.420</td>
</tr>
<tr>
<td>R09</td>
<td>0.560</td>
<td>0.364</td>
<td>0.354</td>
<td>0.408</td>
<td>0.395</td>
</tr>
<tr>
<td>R10</td>
<td>0.629</td>
<td>0.471</td>
<td>0.461</td>
<td>0.513</td>
<td>0.501</td>
</tr>
</tbody>
</table>

Table 3. The aH, aFAR, and WSI for the different models and ensemble forecasts for the October to December 2022 testing period.

<table>
<thead>
<tr>
<th>Model</th>
<th>aH</th>
<th>aFAR</th>
<th>WSI</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMA</td>
<td>86.96</td>
<td>14.80</td>
<td>25.49</td>
</tr>
<tr>
<td>MLR</td>
<td>69.91</td>
<td>32.89</td>
<td>22.50</td>
</tr>
<tr>
<td>RR</td>
<td>69.91</td>
<td>32.89</td>
<td>22.50</td>
</tr>
<tr>
<td>QR</td>
<td>69.91</td>
<td>32.89</td>
<td>22.50</td>
</tr>
<tr>
<td>WR</td>
<td>69.91</td>
<td>32.89</td>
<td>22.50</td>
</tr>
</tbody>
</table>

4 Conclusions

In this study, we utilized five operational experimental aerosol and air quality models, including NASA GEOS, NRL NAAPS, NOAA GEFS, NOAA HRRR, and NAQFC, to generate a real-time ensemble forecast for extreme air quality events such as wildfires. The study period spans across the entire year of 2022. The forecast results of these five models exhibit notable differences, particularly in the major fire regions, which suggest a significant level of uncertainty in the wildfire air quality prediction.

We evaluated the multimodel average (MMA) with each individual model and found that the MMA consistently outperformed individual models in EPA Regions 1, 3, 4, and 10, and in most cases in Regions 5, 6, 7, and 9, whereas some individual models performed better than MMA in Regions 2 and 8. Overall, the MMA performs better than individual models, indicating that ensemble forecasts can effectively reduce forecast uncertainty. In terms of the PM$_{2.5}$ exceedance forecast, MMA achieved the highest weighted success index, which indicates that the ensemble forecast performed better than the individual models in the aspect of extreme events forecast. The ensemble forecast using the MMA effectively reduces forecast uncertainty by cancelling off the systematic errors of individual models in most regions, resulting in a better overall performance. However, the MMA still tends to overestimate in Regions 4, 6, and 10, and underestimate in Regions 5, 8, and 9.

Next, we explored several weighted ensemble methods to further improve the ensemble forecast. Among these methods, we first tested multilinear regression (MLR). Compared to the unweighted ensemble (MMA), the MLR weighted ensemble significantly reduced the forecast
bias by 25% to 53% in the major fire regions. Additionally, the MLR method increased the hit rate by 17%, reduced the false alarm rate by 72%, and raised the WSI by 5% as compared to MMA. These results demonstrate the superiority of the weighted ensemble over the unweighted ensemble.

The second weighted ensemble method we employed was ridge regression (RR), which can alleviate overfitting issues. The results of RR were similar to those of MLR. In the western coast Regions 9 and 10, RR exhibited a lower bias than MLR, indicating that RR might help address the overfitting problem in these large wildfire regions.

Traditional regression methods estimate the conditional mean of the response variable across values of the predictor variables. While this approach works well for forecasting general events, it is not suitable for extreme events. To address this issue, we introduce two additional regression models -- quantile regression (QR) and weighted regression (WR) -- to improve the air quality forecast of extreme events. QR allows the ensemble model to predict more high PM$_{2.5}$ events compared to MLR and MMA, reducing underestimation in large fire regions 8, 9, and 10. However, QR may overestimate the pollution level when the actual pollution level is not high. QR has a much higher hit rate compared to MMA and MLR, with a 55% improvement over MMA and a 33% improvement over MLR. However, its false alarm rate is also higher, reaching 32.89%. QR has the highest WSI among all models, including individual models and ensemble forecasts.

In the case of WR, we assigned more weight to the instances where PM$_{2.5}$ concentration exceeded 20 μg/m$^3$. As a result, WR performed better than MMA and MLR for high PM$_{2.5}$ cases, with fewer false high values compared to QR. While WR had a higher area hit rate than MMA, MLR, and RR, it was lower than QR. Its false alarm rate was slightly higher than MLR and RR, but much lower than MMA and QR. WR had the second-highest WSI among all models.

In summary, there are great uncertainties in wildfire air quality forecasts. Ensemble forecasts can reduce wildfire air quality forecast uncertainties, and overperform individual models most of the time. The weighted ensemble works better than the unweighted ensemble. Quantile regression and weighted regression help forecast extreme cases. Overall, the real-time air quality forecast system could be a useful tool for decision making.

Acknowledgments
This study is financially supported by NASA Health and Air Quality Program, NOAA Weather Program Office, and Disaster Supplemental Program. We thank NASA, NOAA, and NRL for providing the model prediction data used for constructing the ensemble forecast. The views expressed are those of the authors and are not reflective of the federal agencies (NOAA, NASA, NRL, etc.) or institutions.

Open Research
The model data can be downloaded from: http://air.csiss.gmu.edu/yli/paper_data

References


Lucchesi, R., Merkova, D., Nielsen, J. E., Partyka, G., Pawson, S., Putman, W.,
Rienecker, M., Schubert, S. D., Sienkiewicz, M., and Zhao, B.: The Modern-Era
Retrospective Analysis for Research and Applications, Version 2 (MERRA-2), J.

Impact of inherent meteorology uncertainty on air quality model predictions, J. Geophys.

Predictions of 2009’s Tropical Cyclones Initialized with an Ensemble Kalman Filter.
Herrera, M., Alfonso P. Ramallo-González, Matthew Eames, Aida A. Ferreira, David A. Coley,
Creating extreme weather time series through a quantile regression ensemble,
Environmental Modelling & Software, Volume 110, 2018, Pages 28-37, ISSN 1364-8152, https://doi.org/10.1016/j.envsoft.2018.03.007

Hogan, T., Liu, M., Ridout, J., Peng, M., Whitcomb, T., Ruston, B., Reynolds, C., Eckermann,
S., Moskaitis, J., Baker, N., McCormack, J., Viner, K., McLay, J., Flatau, M., Xu, L.,
assimilation by filtering, correction, and aggregation of MODIS Collection 5 optical
https://doi.org/10.5194/amt-4-379-2011

Tong, D.; Tang, Y. Updating and Evaluating Anthropogenic Emissions for NOAA’s
Global Ensemble Forecast Systems for Aerosols (GEFS-Aerosols): Application of an

Global Mortality Attributable to Smoke from Landscape Fires. Environ Health Perspect.
Wildfire smoke injection heights: Two perspectives from space, Geophys. Res. Lett., 35,
METEOROL CLIMATOL, 46, 549–555. https://doi.org/10.1175/JAM2479.1
Khan, N., Shamsuddin Shahid, Liew Juneng, Kamal Ahmed, Tarmizi Ismail, Nadeem Nawaz,
Prediction of heat waves in Pakistan using quantile regression forests, Atmospheric
Research, Volume 221, 2019, Pages 1-11, ISSN 0169-8095,
https://doi.org/10.1016/j.atmosres.2019.01.024.
978-0-521-60827-5.
O’Neill, Susan & Xian, Peng & Flemming, Johannes & Cope, Martin & Baklanov, Alexander & Larkin, Narasimhan & Vaughan, Joseph & Tong, Daniel & Howard, Rosie & Stull, Roland & Davignon, Didier & Ahmadov, Ravan & Odman, Mehmet & Innis, John & Azzi, Merched & Gan, Christopher & Pavlovic, Radenko & Chew, Boon Ning & Reid,


