Bulletin of the American Meteorological Society

Multi-agency Wildfire Air Quality Ensemble Forecast in the United States: Toward Community Consensus of Early Warning

--Manuscript Draft--

Manuscript Number:	BAMS-D-23-0208
Full Title:	Multi-agency Wildfire Air Quality Ensemble Forecast in the United States: Toward Community Consensus of Early Warning
Article Type:	Article
Order of Authors:	Yunyao Li, Ph.D.
	Daniel Tong
	Peewara Makkaroon
	Timothy DelSole
	Youhua Tang
	Patrick Campbell
	Barry Baker
	Mark Cohen
	Anton Darmenov
	Eric James
	Ravan Ahmadov
	Edward Hyer
	Peng Xian
Manuscript Classifications:	7.012: Ensembles; 7.020: Forecasting; 12.028: Aerosols/particulates; 12.036: Air quality; 12.144: Wildfires; 12.068: Biomass burning
Abstract:	Wildfires pose increasing risks to human health and properties in North America. Due to large uncertainties in fire emission, transport, and chemical transformation, it remains challenging to accurately predict air quality during wildfire events, hindering our collective capability to issue effective early warnings to protect public health and welfare. Here we present a new real-time Hazardous Air Quality Ensemble System (HAQES) by leveraging various wildfire smoke forecasts from three U.S. federal agencies (NOAA, NASA, and Navy). Compared to individual models, the HAQES ensemble forecast significantly enhances forecast accuracy. To further enhance forecasting performance, a weighted ensemble forecast approach was introduced and tested. Compared to the unweighted ensemble mean, the multilinear regression weighted ensemble reduced fractional bias by 34% in the major fire regions, false alarm rate by 72%, and increased hit rate by 17%. Finally, we improved the weighted ensemble using quantile regression and weighted regression methods to enhance the forecast of extreme air quality events. The advanced weighted ensemble increased the PM2.5exceedance hit rate by 55% compared to the ensemble mean. Our findings provide insights into the development of advanced ensemble forecast methods for wildfire air quality, offering a practical way to enhance decision-making support to protect public health.

Multi-Agency Ensemble Forecast of Wildfire Air Quality in the United States: Toward Community Consensus of Early Warning

Yunyao Li,[a,b] Daniel Tong,[a,b] Peewara Makkaroon,[a] Timothy DelSole,[a] Youhua Tang,[b,c] Patrick Campbell,[b,c] Barry Baker,[c] Mark Cohen,[c] Anton Darmenov,[d] Ravan Ahmadov,[e] Eric James,[e] Edward Hyer,[f] Peng Xian [f].

[a] *Department of Atmospheric, Oceanic and Earth Sciences, George Mason University, Fairfax, VA, USA*

[b] *Center for Spatial Information Science and Systems, George Mason University, Fairfax, VA, USA*

[c] *Air Resources Laboratory, National Oceanic and Atmospheric Administration, College Park, MD 20740, USA*

[d] *National Aeronautics and Space Administration Goddard Space Flight Center, Greenbelt, USA*

[e] *Cooperative Institute for Research in Environmental Sciences, University of Colorado at Boulder, Boulder, CO, USA*

[f] *Marine Meteorology Division, Naval Research Laboratory, Monterey, CA, USA*

Corresponding author: Yunyao Li, yli74@gmu.edu, Daniel Tong, qtong@gmu.edu

ABSTRACT

Wildfires pose increasing risks to human health and properties in North America. Due to large uncertainties in fire emission, transport, and chemical transformation, it remains challenging to accurately predict air quality during wildfire events, hindering our collective capability to issue effective early warnings to protect public health and welfare. Here we present a new real-time Hazardous Air Quality Ensemble System (HAQES) by leveraging various wildfire smoke forecasts from three U.S. federal agencies (NOAA, NASA, and Navy). Compared to individual models, the HAQES ensemble forecast significantly enhances forecast accuracy. To further enhance forecasting performance, a weighted ensemble forecast approach was introduced and tested. Compared to the unweighted ensemble mean, the multilinear regression weighted ensemble reduced fractional bias by 34% in the major fire regions, false alarm rate by 72%, and increased hit rate by 17%. Finally, we improved the weighted ensemble using quantile regression and weighted regression methods to enhance the forecast of extreme air quality events. The advanced weighted ensemble increased the $PM_{2.5}$ exceedance hit rate by 55% compared to the ensemble mean. Our findings provide insights into the development of advanced ensemble forecast methods for wildfire air quality, offering a practical way to enhance decision-making support to protect public health.

SIGNIFICANCE STATEMENT

Wildfires are a growing threat to health and safety in North America. Accurately predicting air quality during these events is crucial but challenging. In response, we've developed the real-time Hazardous Air Quality Ensemble System (HAQES), by combining forecasts from three U.S. federal agencies (NOAA, NASA, and Navy). HAQES significantly improves accuracy compared to individual models. Moreover, we further improve the wildfire air quality forecast by introducing the weighted ensemble method. The weighted ensemble reduced bias by 34% and false alarms by 72%, while increasing hit rates by 55%. HAQES advances our ability to protect public health during wildfire events.

CAPSULE (BAMS ONLY)

We built a real-time multi-model Hazardous Air Quality Ensemble System leveraging operational/research forecasts from U.S. federal agencies and developed weighted ensemble methods to enhance wildfire air quality forecasts.

1. Introduction

Wildfires are a significant contributor to atmospheric aerosols and trace gases, causing hazardous air quality and adverse health effects. Research has established links between wildfire smoke exposure and all-cause mortality, as well as respiratory health issues (Cascio, 2018). The global average mortality attributable to landscape fire smoke exposure was estimated to be 339,000 deaths annually (Johnston et al., 2012).

Air quality forecast during wildfire events is crucial for public health management and emergency response, including early warnings, but it remains a challenging task due to uncertainties in fire emissions (Pan et al., 2020), plume rise calculations (Ye et al., 2021; Li et al., 2023), and other model inputs/processes (Delle Monache and Stull, 2003).

Ensemble forecasting techniques have been increasingly used to improve the predictability of extreme air quality episodes. 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. Li et al. (2020) used an ensemble forecast to predict surface $PM_{2.5}$ during the 2018 California Camp Fire event using the National Oceanic and Atmospheric Administration (NOAA) Hybrid Single-Particle Lagrangian Integrated Trajectory (HYSPLIT) dispersion model with different emissions, plume heights, and model setups. Makkaroon et al. (2023) successfully demonstrated a multi-model ensemble forecast system that effectively simulated the 2020 western US "Gigafire", with the ensemble mean outperforming individual models. These studies highlight the potential of ensemble forecasting to improve the predictability of wildfire air quality.

While multi-model ensemble often outperforms single-model forecasts, some challenges remain. The ensemble mean does not work best all the time (Xian et al., 2019; Makkaroon et al., 2023). For instance, insufficient diversity among models in the multi-model ensemble can limit the ability of the ensemble to capture the full uncertainties and variability tied to different inputs and assumptions. Moreover, if individual models in the ensemble are biased, the ensemble itself may exhibit systematic bias (DelSole et al., 2016).

This study presents a new Hazardous Air Quality Ensemble System (HAQES) over the Contiguous United States (CONUS) by leveraging real-time forecasts from three U.S. federal agencies (NOAA, NASA, and Navy). We applied a weighted ensemble forecast approach to

enhance skill and further improved it by incorporating quantile regression, weighted regression methods to enhance extreme air quality forecasts, and ridge regression to address overfitting concerns. We also introduced a combination of random walk and categoric metrics to assess the performance of the ensemble and individual models against AirNow observations for the year 2022.

2. Materials and Methods

a. Fires in 2022

This paper focuses on the year 2022 when 66,255 fires (12th most since 2001) burned 7,534,403 acres (11th most), as reported by the National Interagency Fire Center. Figure 1 displays the annual and monthly total fire radiative energy (FRE) from Global Biomass Burning Emissions Product (GBBEPx; Zhang et al., 2019), which is highly correlated with fire emissions (Wooster et al., 2005), across the 10 U.S. Environmental Protection Agency (EPA) regions for 2022. In the eastern U.S., biomass-burning emissions were concentrated in the southeastern states (Region 4). Although the Southeast fires affected a large area, the total FRE was not as high as that of the western wildfires. Region 4's peak fire period was in March, releasing 6,281 TJ of fire energy in one month. In the central U.S., fire emissions were primarily located in Regions 6 and 7. Central U.S. fires peaked in spring (April-May), releasing 41,634 TJ of fire energy 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 occurred in September (33,631 TJ), followed by May (30,330 TJ).

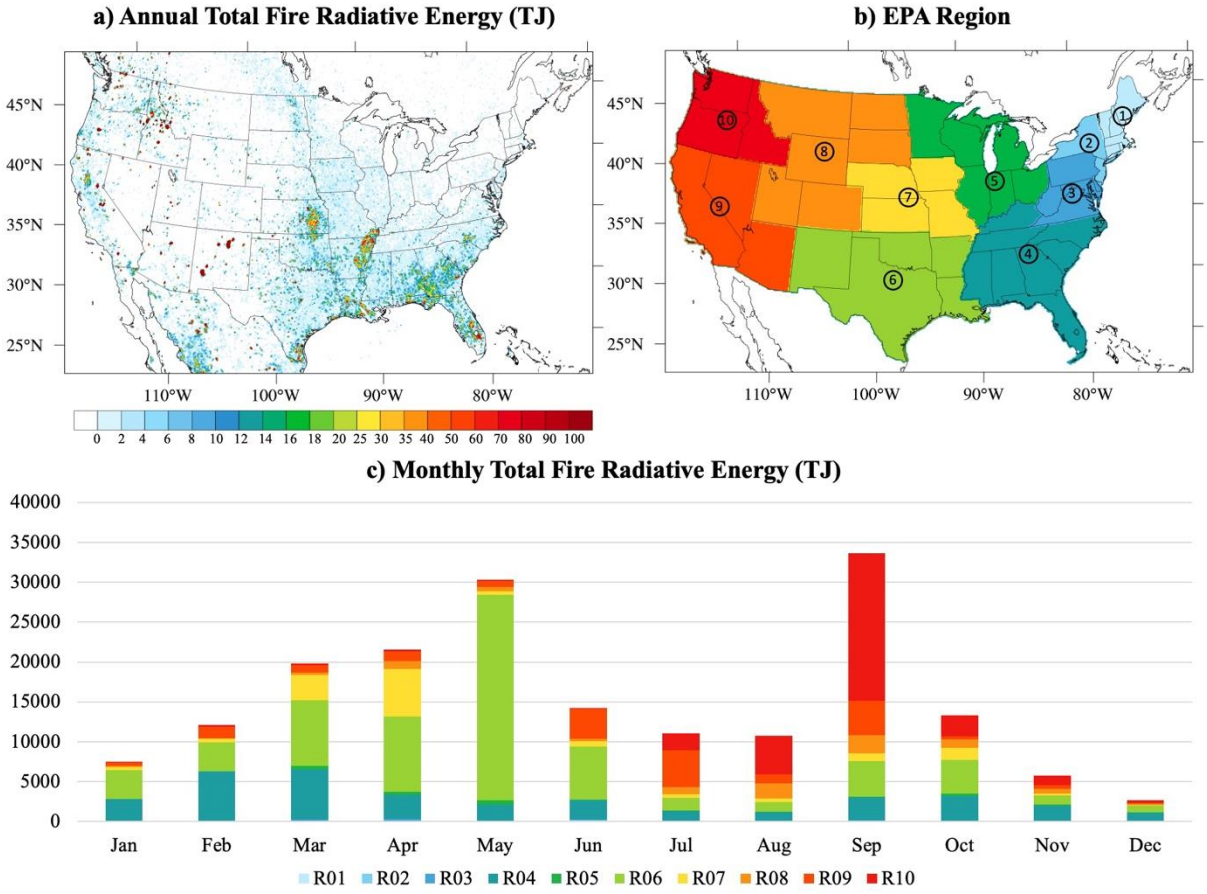

Fig. 1. The annual (a) and monthly (c) total fire radiative energy across the 10 U.S. EPA regions (b) for 2022.

b. Description of Ensemble Members

The air quality forecast ensemble in this study was developed using both regional and global chemical transport models, including the NOAA High-Resolution Rapid Refresh-Smoke (HRRR-Smoke), Global Ensemble Forecast System Aerosols (GEFS-Aerosols), National Air Quality Forecasting Capability (NAQFC), the NASA Goddard Earth Observing System (GEOS), and the Naval Research Laboratory (NRL) 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 resolution. The ensemble exploits the strengths of these widely different models to improve forecasting accuracy. These models encompass a wide range of emission datasets and 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.

1) HRRR-SMOKE

HRRR-Smoke (Ahmadov, et al., 2017; Dowell et al., 2022) is an operational real-time three-dimensional coupled weather-smoke forecast model operating at a 3 km spatial resolution over the Continental United States (CONUS) domain, maintained by NOAA National Centers for Environmental Prediction (NCEP). The HRRR Data-Assimilation System provides initial conditions and a background ensemble for meteorological data assimilation. HRRR-Smoke ingests the satellite fire radiative power data (FRP) from the Suomi-NPP (National Polar-orbiting Partnership), NOAA-20, and Moderate Resolution Imaging Spectroradiometer (MODIS) Terra/Aqua satellites to estimate wildfire smoke emissions. Since HRRR-Smoke is designed to forecast $PM_{2.5}$ where smoke is a dominant pollution source, it does not include any non-fire emissions (e.g., anthropogenic emissions) and gas/aerosol chemistry.

2) GEFS-AEROSOL

GEFS-Aerosols (Zhang et al., 2022) is a global atmospheric composition model developed by the NCEP in collaboration with the NOAA Global Systems Laboratory, Chemical Sciences Laboratory, and Air Resources Laboratory. It integrates Finite Volume Cubed Sphere (FV3)-based Global Forecast System (GFS) version 15 meteorology and WRF-Chem's atmospheric aerosol chemistry. The Aerosol module is based on the NASA Goddard Chemistry Aerosol Radiation and Transport model (GOCART) (Chin et al., 2002) with both fire emission and anthropogenic emission. The biomass-burning emission is from GBBEPx. Smoke plume rise is calculated using a one-dimension 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 0.25° × 0.25° resolution.

3) NAQFC

NOAA's operational NAQFC uses CMAQ version 5.3.1 driven by NOAA's latest operational FV3-GFSv16 meteorology at the horizontal spatial resolution of 12 km with 35 vertical layers (Campbell et al., 2022). The chemical gaseous boundary conditions are based on static, global GEOS-Chem simulations, while aerosol boundary conditions are dynamically updated from NOAA’s operational GEFS-Aerosols model. NAQFC employs GBBEPx for biomass-burning emissions. The model uses the Briggs (1969) plume rise

algorithm to compute wildfire smoke plumes. It also includes anthropogenic emissions and biogenic emissions.

4) GEOS

The GEOS (Gelaro et al., 2017) system was developed by NASA's Global Modeling and Assimilation Office. This study used the GEOS Forward Processing system (GEOS-FP, version 5.27.1) at a 0.25° × 0.3125° spatial resolution, which generates analyses, assimilation products, and ten-day forecasts in near-real time. GEOS-FP is built around the GEOS Atmospheric General Circulation Model, 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 and are simulated with the GOCART (Chin et al., 2002). Fire emissions come from the Quick Fire Emissions Dataset (QFED; Darmenov and da Silva, 2015) and leverage low-latency MODIS fire locations and FRP (Collection 6) data. Emissions from fires are distributed in the Planetary Boundary Layer (PBL). The model also includes anthropogenic and biogenic emissions.

5) NAAPS

NAAPS (Lynch et al., 2016) is developed at NRL Marine Meteorology Division 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° × 0.333°. The current NAAPS is driven by global meteorological fields from the Navy Global Environmental Model (NAVGEM; 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).

c. Description of Observations

The hourly ground $PM_{2.5}$ observations from the U.S. EPA AirNow network for the year 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 the EPA. It contains

177 air quality data for more than 500 cities across the U.S. (total of 1156 sites), as well as for
178 Canada and Mexico.

179 *d. Ensemble design*

180 In this study, we examined five techniques for creating ensembles categorized into two
181 groups: unweighted and weighted ensemble approaches. Unweighted ensemble employed
182 multi-model average (MMA) to merge predictions from multiple models into one
183 consolidated forecast, while weighted ensemble assigned different weights (β) to member
184 models (M_j):

185 $$\widehat{M} = \sum_{j=1}^{S} \beta_j M_j + \beta_0 \quad (1)$$

186 where S represents the total number of models which is 5 in this study, and β_0 is the intercept.
187 To determine the weights, the data for the year 2022 are grouped into training and testing
188 sets. Since wildfires can last for weeks, to ensure the independence of the training and testing
189 data, we did not select the training data randomly. Instead, we used the first 9 months of data
190 as the training set and the final 3 months as the testing set. Due to computational limitations
191 (space and time), we were only able to analyze one year of data, which may lead to
192 variability in the calculated weights for each model. However, the purpose of this paper is to
193 introduce and test various weighted ensemble approaches for air quality forecasting. Longer
194 training and testing periods are required before implementing a weighted ensemble in
195 operational forecasting, to thoroughly investigate its performance and determine the optimal
196 weights for each model.

197 We experimented with four regression methods to determine these weights: Multi-linear
198 Regression, Ridge Regression, Quantile Regression, and Weighted Regression.

199 1) MULTI-LINEAR REGRESSION (MLR)

200 MLR calculates the weights for each model by minimizing the error between the
201 observation (O) and the weighted multimodel prediction:

202 $$\hat{\beta}^{\mathrm{MLR}} = arg\,\min_{\beta}\left(\sum_{i=1}^{N}\left(O_i - \beta_0 - \sum_{j=1}^{S}\beta_j\, M_{ij}\right)^2\right) \quad (2)$$

where N is the total number of observations.

2) Ridge Regression (RR)

The ridge regression (Hoerl and Kennard, 1970) is a technique used to reduce overfitting issues in MLR, which is a common problem in statistical modeling and machine learning. RR adds a penalty term to the 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:

$$\hat{\beta}^{RR} = \arg\min_{\beta}\left(\sum_{i=1}^{N}\left(O_i - \beta_0 - \sum_{j=1}^{S}\beta_j M_{ij}\right)^2 + \lambda\sum_{j=1}^{S}\beta_j^2\right) \quad (3)$$

where λ is the ridge parameter. The first 20 days in each month are used to train the data using Eq (3), and the last 10 days are used to find the best λ. We tested the value of λ from 1 to 1000. Ridge regression can produce a more robust and stable model, especially when the number of predictors is large, and the predictors are nearly collinear, which occurs often in multi-model forecasting. It has been found to be useful in climate ensemble studies (DelSole et al., 2007).

3) Quantile Regression (QR)

MLR and RR estimate the conditional mean of the forecast and tend to favor the mean state, which is suitable for general cases, but not for extreme events. To address this, we employ QR to enhance extreme air quality ensemble forecasting (Koenker and Bassett, 1978). QR is an approach like traditional linear regression but with quantile-dependent regression coefficients:

$$\widehat{M}_{QR} = \sum_{j=1}^{S}\beta_{j,q}M_j + \beta_{0,q} \quad (4)$$

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:

$$\hat{\beta}^{QR} = \arg\min_{\beta}\left(\sum_{j:M\geq M_q} q\left|O_i - \beta_{0,q} - \sum_{j=1}^{S}\beta_{j,q}M_{i,j}\right|\right.$$

$$\left. + \sum_{j:M<M_q}(1-q)\left|O_i - \beta_{0,q} - \sum_{j=1}^{S}\beta_{j,q}M_{i,j}\right|\right) \quad (5)$$

4) WEIGHTED REGRESSION (WR)

WR is another statistical method addressing the issue of extreme events. WR 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:

$$\hat{\beta}^{WR} = arg\min_{\beta}\left(\sum_{i=1}^{N} W_i\left(O_i - \beta_0 - \sum_{j=1}^{S}\beta_j M_{i,j}\right)^2\right) \quad (6)$$

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 μg/m^3 (80% of the total observations, based on Kang et al. (2007), the basis for calculating the weighted success index for extreme forecast), and a weight of 1 to other points, which gives more importance to polluted days:

$$W_i = \begin{cases} 10, & if\ O_i > 20\ \mu g/m^2 \\ 1, & otherwise \end{cases} \quad (7)$$

e. Evaluation method

1) RANDOM WALK

We employ the DelSole and Tippett (2016) random walk method to evaluate the performance of both ensemble and individual models. When comparing forecasts A and B for N times, a positive step is taken if A outperforms B, and a negative step if otherwise. Let K represent the number of times that forecast A outperforms forecast B. The net distance (d; forecast score) traveled by the random walk is:

$$d_N = K - (N - K) = 2K - N \quad (8)$$

Fractional bias (FB, Appendix A) is used to determine the more skillful forecast for each event. A significance test (K_α, Appendix B) is conducted to show if A is significantly better ($K > K_\alpha$) or worse ($K < N - K_\alpha$) than B.

2) CATEGORICAL METRICS

Standard metrics like fractional bias 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 US EPA National Ambient Air Quality Standards (NAAQS) 24-hour $PM_{2.5}$ exceedance events (>35 $\mu g/m^3$; U.S. EPA, 2020). Here, we used three categorical metrics (0-100%) described by Kang et al. (2007):

(1) Area hit rate (aH) - indicates match between forecasted and observed poor air quality exceedances. Higher aH implies a more reliable model.

(2) Area false alarm rate (aFAR) - measures incorrect predictions of poor air quality. Lower aFAR implies a more reliable model.

(3) Weighted success index (WSI) - considers hits, false alarms, and missed exceedance forecasts. A higher WSI suggests a more reliable model.

The equations for these metrics are shown in Appendix C.

3. Results

This section begins with evaluating the performance of the unweighted ensemble forecast compared to each individual model (referred to as model-1 through 5; The purpose of this study is to assess ensemble forecast skills rather than delving into the performance of individual models. We intentionally rearranged the order of these models and renamed them to models 1-5 to avoid focusing on specific model performance.). Secondly, we compare the performance of the unweighted ensemble with that from different weighted ensemble methods.

a. Comparison of MMA with individual models

The annual mean surface $PM_{2.5}$ concentration (Fig. 2) predicted by models 1 to 5 and the MMA are compared to the AirNow observations. The results from different models varied substantially, highlighting the significant uncertainty in wildfire air quality forecasts. Models 1, 2, and 4 overestimate $PM_{2.5}$ in the Southeast and Northwest, where models 3 and 5

underestimate it. The ensemble mean balanced these overestimations and underestimations and is closer to the observations.

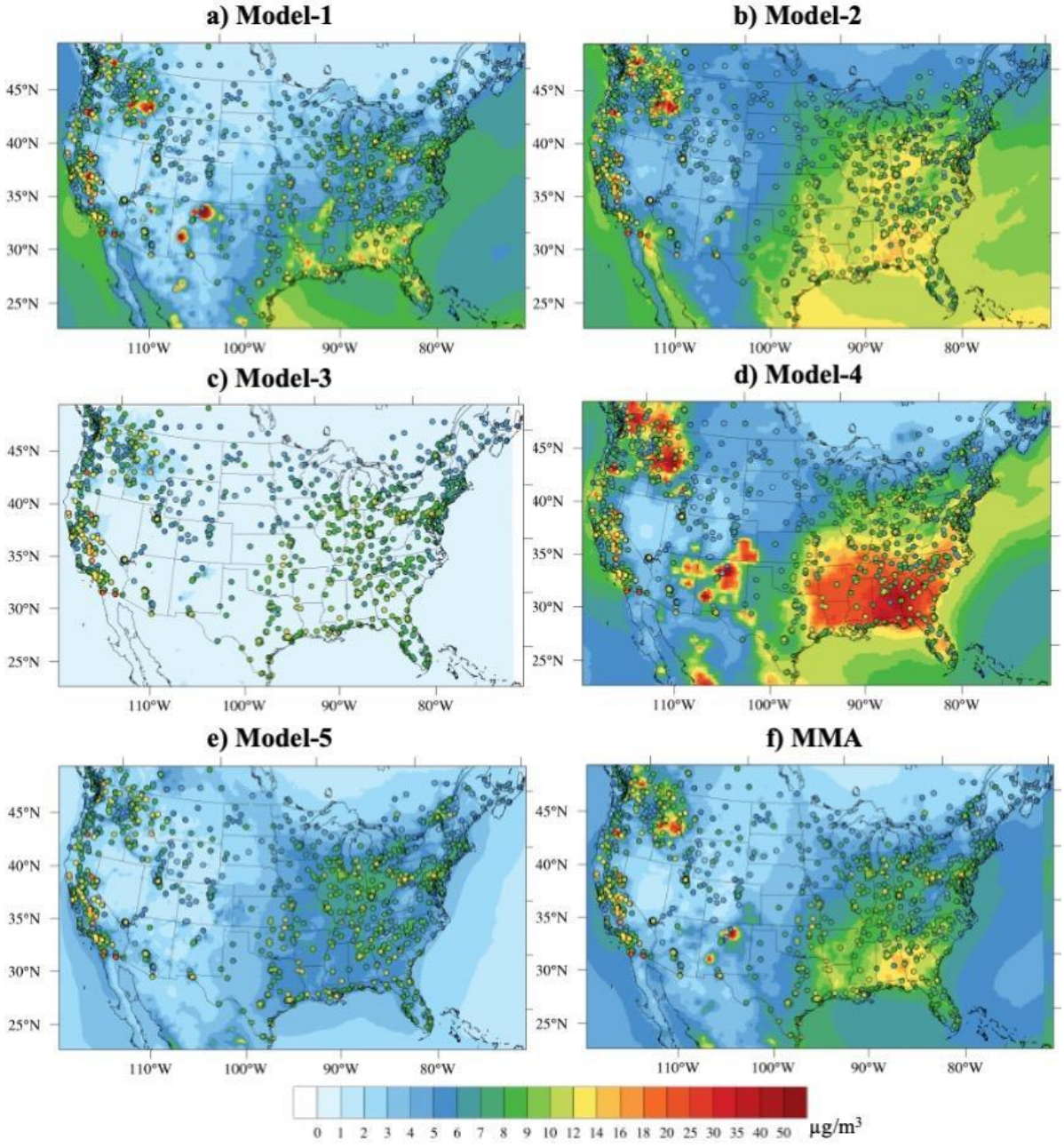

Fig. 2. Annual mean surface $PM_{2.5}$ concentration (contour) predicted by models 1 to 5 and MMA and observed by the AirNow network (colored circles) for the year of 2022.

We compared the MMA with each individual model using the random walk method for major fire regions (EPA region 4, 6, 7, 9, and 10 from Fig. 1c; Fig. 3). Figure 3 shows the relative forecast score of individual models compared to MMA. A negative value on day n indicates that the overall performance of MMA surpasses that of the individual model from

day 1 to day n; the negative trend observed from day n_1 to n_2 signifies that the MMA consistently outperforms the individual model between n_1 and n_2, and vice versa. In regions 4 and 10, the consistent downward trend of the random walk scores implies that MMA consistently outperforms individual models. In regions 6, 7, and 9, the scores are mostly negative with some transient positive scores in early 2022 as well as some positive tendency at the end of the year (because of the underestimate of the anthropogenic emission), indicating that MMA performs better than each model most of the time, especially in the wildfire season (Summer and Fall). The MMA improves air quality forecasting because it balances the model bias of the five individual models (Fig. S1). Overall, the MMA outperforms individual models, demonstrating that ensemble forecasts can effectively reduce forecast uncertainty.

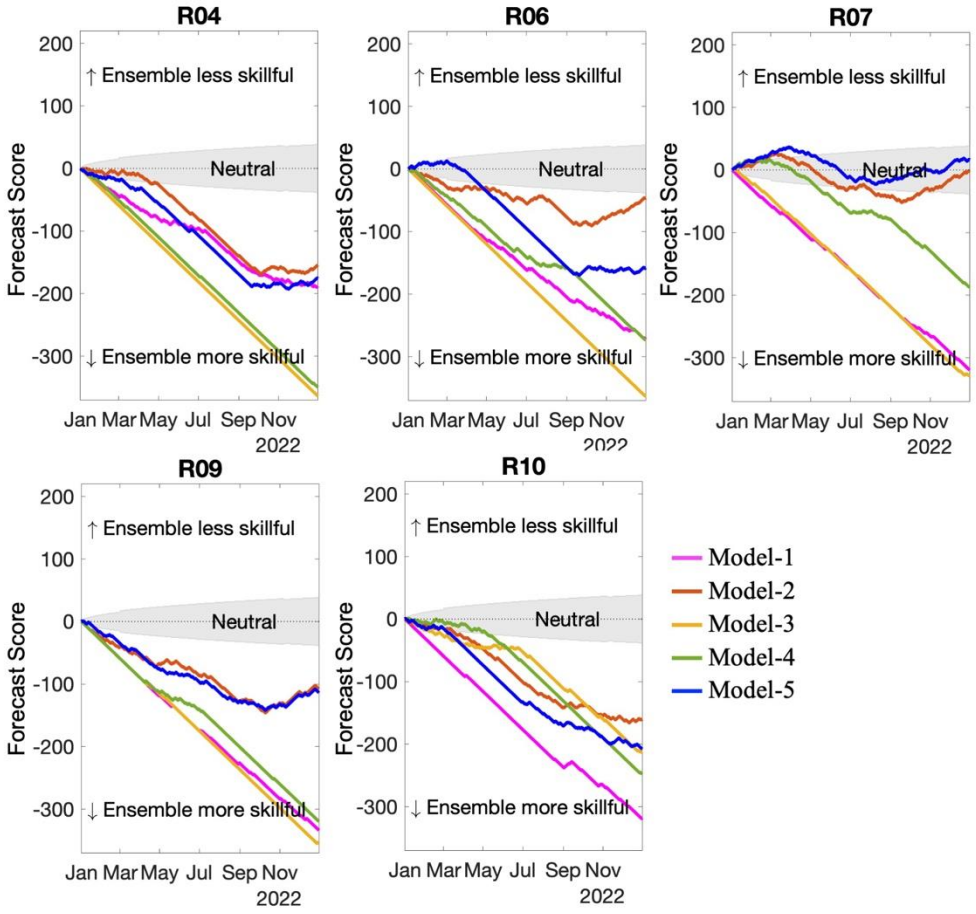

Fig. 3. Compare MMA to individual models using the random walk method for major fire regions.

To evaluate the forecasting ability of extreme events by individual models and MMA, we calculated the area hit rate, area false alarm rate, and weighted success index for the year

2022 (Table 1). The MMA obtains the highest WSI, third highest hit rate, and the second lowest false alarm rate. Model 4 excels in hit rate but has the highest false alarms. Model 3 has the lowest false alarm rate, but also the lowest hit rate. Overall, the MMA ensemble works better than the individual models in extreme events air quality forecast, consistent with prior research (Li et al., 2020; Makkaroon et al., 2023).

	Model-1	**Model-2**	**Model-3**	**Model-4**	**Model-5**	**MMA**
aH	26.98	41.12	14.68	**48.12**	<u>12.42</u>	37.44
aFAR	83.29	79.70	**29.77**	<u>93.10</u>	84.17	77.09
WSI	13.06	20.10	16.47	<u>6.98</u>	13.82	**20.68**

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 year 2022. The best results are highlighted in bold, while the worst results are underlined.

b. Weighted ensemble

MMA improved air quality forecasting, but there is still room for further improvement. Therefore, we explored various weighted ensemble approaches to further enhance forecasting performance. As explained in section 2.d, the initial 9 months are utilized for weight calculation, while the subsequent 3 months serve as the testing data, which will be assessed in this section. The first weighted ensemble approach we tested is MLR. Compared to the MMA, MLR reduces the fractional bias by 34%, increases the hit rate by 17%, significantly reduces the false alarm rate by 72%, and increases the WSI by 5% (Table 2) and is closer to the observations (Fig. 4). These results demonstrate that the weighted ensemble outperforms the unweighted ensemble.

	M-1	**M-2**	**M-3**	**M-4**	**M-5**	**MMA**	**MLR**	**RR**	**QR**	**WR**
FB	0.60	0.49	<u>1.87</u>	0.88	0.50	0.50	**0.33**	0.34	0.41	0.35
aH	42.09	76.87	<u>8.52</u>	65.22	61.04	56.00	65.39	61.04	**86.96**	69.91
aFAR	<u>57.24</u>	31.25	**0**	51.64	32.21	29.11	8.14	6.17	32.89	14.80

WSI	5.09	17.03	1.18	12.04	18.52	19.16	20.15	16.92	**25.49**	22.50

Table 2. The Fractional bias (FB), aH, aFAR, and WSI for the different models (Model-1 to Model-5, M1-M5), MMA, and four weighted ensemble forecasts (MLR, RR, QR, and WR) for the October to December 2022 testing period (bold represents the best results, underline represents the worst results).

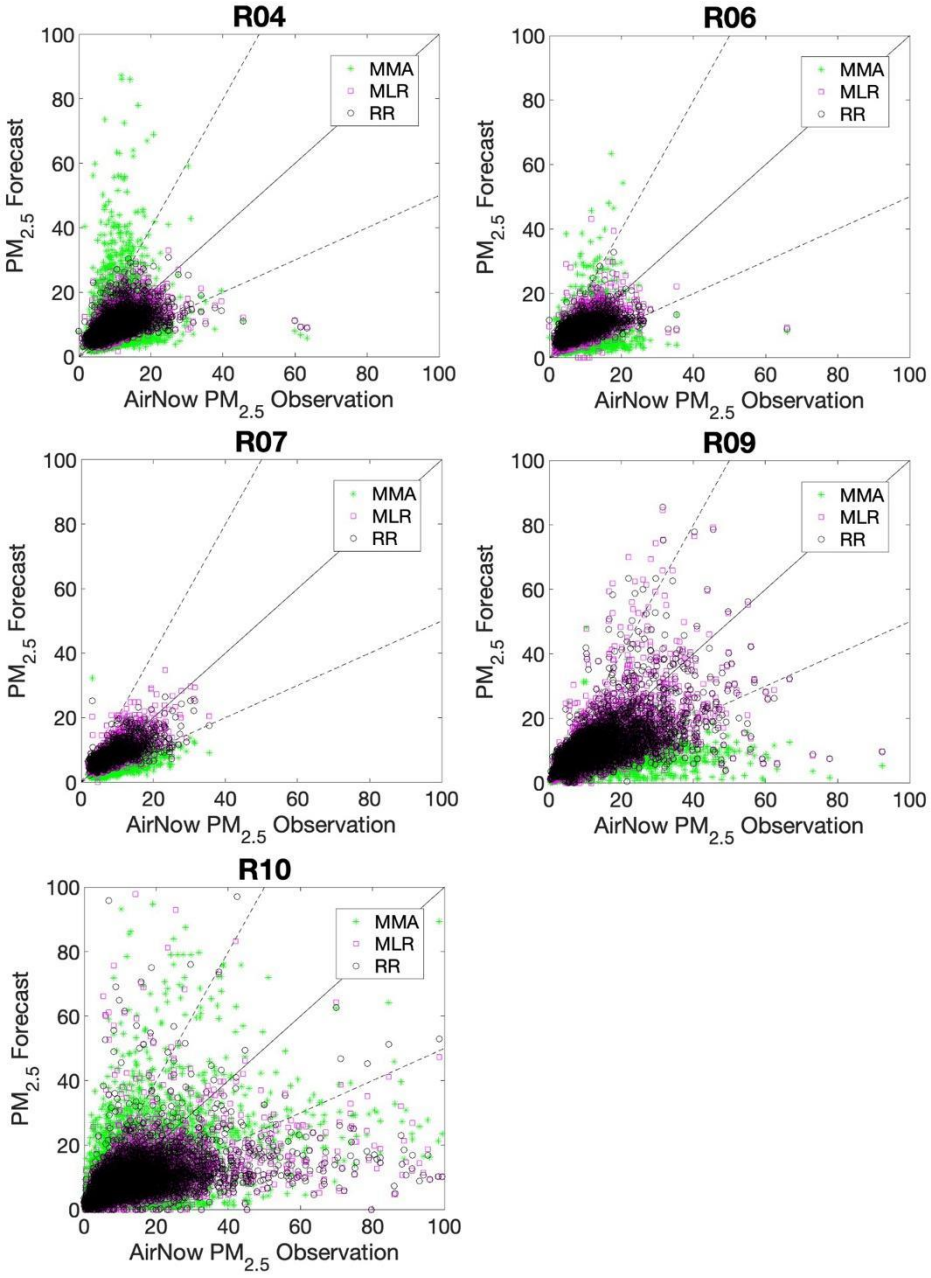

Fig. 4. Scatter plots between predicted and observed $PM_{2.5}$ for MMA (green), MLR (magenta), and RR (black) for five fire-prone EPA regions. The solid black line represents

the 1:1 ratio line for the observations and forecasts, while the dashed black lines represent the 1:2 and 2:1 ratio lines.

The performance of RR is generally comparable to that of MLR (Fig. 4). RR has a slightly lower hit rate, lower false alarm rate, and lower weighted success index (Table 2) compared to MLR. Employing RR to mitigate the overfitting concern of MLR doesn't notably enhance model performance. This could be attributed to the modest number of models, so the data is not too noisy. Previous studies found that RR can produce a more robust and stable model when the number of predictors is large and the data is noisy (DelSole et al., 2007; Pena and van den Dool, 2008).

MMA, MLR, and RR all tend to underestimate the $PM_{2.5}$ exceedance events (Fig. 4), particularly in the Western Coast with high wildfire emissions (R9 and 10). Therefore, we applied QR to enhance predictions of extreme cases. QR enables the ensemble model to predict more polluted events than MLR and MMA (Table 2). QR has a much higher hit rate, which is about 55% higher than the MMA and 33% higher than the MLR. However, sometimes QR overestimates the pollution level when the actual pollution level is not high. Its false alarm rate reaches 32.89%. QR has the highest WSI among all models, including individual models and ensemble forecasts. The fractional bias of QR is higher than that of MLR and RR but still 18% lower than that of MMA.

Another approach to improve the ensemble forecast's ability to predict extreme cases is WR. WR improved the forecast for $PM_{2.5}$ exceedance by increasing the area hit rate by 7% compared to MLR and 25% compared to MMA, respectively. QR focuses on the top 10% of cases, whereas WR assigns greater weight to the top 20% of cases. Consequently, predictions using QR tend to be higher than WR (Fig. S6). QR exhibits more overestimation and less underestimation compared to WR. As reflected in Table 2, the Fractional Bias of QR is 15% higher than WR, the area Hit rate of QR is 17 % higher, and the false alarm rate is also elevated (55%) compared to WR. WR offers a balanced enhancement. WR’s WSI is the second highest which surpasses MMA and all individual models.

4. Conclusions

In this study, we built a new real-time Hazardous Air Quality Ensemble System (HAQES) by leveraging operational and research fire wildfire smoke forecasts from U.S. federal agencies: GEOS from NASA, NAAPS from NRL, and GEFS-Aerosol, HRRR-Smoke, and

NAQFC from NOAA. Automated transfer links have been established between these agencies and GMU. Individual model daily forecast results are automatically transmitted to GMU each day to generate the real-time ensemble forecast results. HAQES significantly enhances forecast accuracy compared to single model forecasts, reducing model bias and increasing the weighted success index for $PM_{2.5}$ exceedances.

To further enhance forecasting performance, we introduced a weighted ensemble forecast using multilinear regression. Compared to the unweighted ensemble mean, the multilinear regression weighted ensemble reduced model bias by 34%, false alarm rate by 72%, and increased hit rate by 17%. We also used ridge regression to reduce the overfitting issue of multilinear regression; however, the ridge regression weighted ensemble is close to multilinear regression weighted ensemble, indicating that the overfitting was not significant in our ensemble system.

Finally, we improved the weighted ensemble using quantile regression and weighted regression to enhance the forecasting capability during extreme air quality events. The advanced weighted ensemble increased the hit rate by 55% for $PM_{2.5}$ exceedance compared to that by the ensemble mean. Our findings provide insights into the development of advanced ensemble forecast methods for wildfire air quality, which offers a practical way to enhance decision-making support through leveraging existing forecasting efforts across federal agencies.

Acknowledgments.

This study is financially supported by NASA Health and Air Quality Program and NOAA Weather Program Office. 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 necessarily reflective of the federal agencies (NOAA, NASA, NRL, etc.) or institutions.

Data Availability Statement.

Here are the link for each model: GEFS: https://ftp.ncep.noaa.gov/data/nccf/com/gens/prod; GEOS: https://portal.nccs.nasa.gov/datashare/gmao/geos-fp/forecast; HRRR:

https://nomads.ncep.noaa.gov/pub/data/nccf/com/hrrr/prod; NAQFC: https://airquality.weather.gov; NAAPS: https://usgodae.org/pub/outgoing/fnmoc/models; HAQES: http://air.csiss.gmu.edu/haqes; AirNow data can be downloaded here: https://files.airnowtech.org/?prefix=airnow/2022/.

APPENDIX

Appendix A: Fractional Bias

Below is the definition of fractional bias:

$$FB_i = 2 \times \frac{|O_i - M_i|}{O_i + M_i} \quad (A1)$$

where O is the AirNow observation, and M is the model forecast.

Appendix B: Significance Test (K_α) for Random Walk

Kα can be approximated as:

$$K_\alpha = \left\lceil \frac{N}{2} - z_{\frac{\alpha}{2}} \sqrt{\frac{N}{4} - \frac{1}{2}} \right\rceil \quad (A2)$$

where z_α is the value for which a standardized Gaussian is exceeded with probability α=5%, and ⌈x⌉ denotes ceiling function that maps x to the smallest integer greater of equal to x.

Appendix C: Categorical Metrics

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 monitor location.

$$aFAR = \left(\frac{Aa}{Aa + Ab}\right) \times 100\% \quad (A3)$$

$$aH = \left(\frac{Ab}{Ab + Ac}\right) \times 100\% \quad (A4)$$

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 aFAR (A3) refers to the percentage of false alarms if a forecasted exceedance is not observed within the designated area. The area hit rate aH (A4) refers to the percentage of hits if a forecasted exceedance is observed within the designated area. The aFAR and aH both range from 0-100%. If a model performs well, the misses (Ac) will be low, and the hits (Ab) will be high, resulting in high aH. In contrast, if a model performs poorly, the false positives (Aa) will be high and the hits (Ab) will be low, resulting in high aFAR.

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

$$WSI = \frac{b + \sum_1^n IP}{a + b + c} \times 100\% \quad (A5)$$

$$IP = \begin{cases} \dfrac{M - fO}{M - fT} & if\ O < T < M < fO \\ \dfrac{O - fM}{O - fT} & if\ M < T < O < fM \end{cases} \quad (A6)$$

where a, b, and c refer to the three scenarios defined above, and n represents the total number of observations. Note the choice of f in A6 is empirical and is based on rules of thumb (Hanna 2006). Analysis of $PM_{2.5}$ results for 2022 has shown that about 80% of the difference between observation and prediction is within a factor of 2; thus, in this study, f is set to 2.

REFERENCES

Ahmadov, and Coauthors, 2017: Using VIIRS Fire Radiative Power data to simulate biomass burning emissions, plume rise and smoke transport in a real-time air quality modeling system. *2017 IEEE International Geoscience and Remote Sensing Symposium*. pp. 2806-2808, doi: 10.1109/IGARSS.2017.8127581.

441 Briggs, G., 1969: Plume rise: A critical review (Technical Report). (p. 81). Springfield, VA:
442 National Technical Information Service.

443 Campbell, P., and Coauthors, 2022: Development and evaluation of an advanced National Air
444 Quality Forecasting Capability using the NOAA Global Forecast System version 16.
445 *Geosci. Model Dev.*, 15(8), 3281–3313. https://doi.org/10.5194/gmd-15-3281-2022

446 Cascio W., 2018: Wildland fire smoke and human health. *Sci Total Environ*. doi:
447 10.1016/j.scitotenv.2017.12.086.

448 Chin, M., and Coauthors, 2002: Tropospheric Aerosol Optical Thickness from the GOCART
449 Model and Comparisons with Satellite and Sun Photometer Measurements. *J. Atmos.*
450 *Sci.*, 59(3), 461–483. https://doi.org/10.1175/1520-
451 0469(2002)059<0461:TAOTFT>2.0.CO;2

452 Darmenov, A. and A. da Silva, 2015: The Quick fire emissions dataset (QFED):
453 documentation of versions 2.1, 2.2 and 2.4. Technical Report Series on Global
454 Modeling and Data Assimilation (NASA/TM–2015-104606, Vol.38), NASA Global
455 Modeling and Assimilation Office,
456 https://ntrs.nasa.gov/api/citations/20180005253/downloads/20180005253.pdf

457 Delle Monache, L., and R. Stull, 2003: An ensemble air-quality forecast over western Europe
458 during an ozone episode. *Atmos. Environ.*, 37(25), 3469–3474.
459 https://doi.org/10.1016/S1352-2310(03)00475-8

460 DelSole, T., and M. Tippett, 2016: Forecast Comparison Based on Random Walks, *Mon.*
461 *Weather Rev.*, 144, 615–626, https://doi.org/10.1175/MWR-D-15-0218.1.

462 DelSole, T., 2007: A Bayesian framework for multimodel regression. *J. Climate*, 20, 2810–
463 2826.

464 Dowell, D., and Coauthors, 2022: The high- resolution rapid refresh (HRRR): An hourly
465 updating convection- allowing forecast model. Part 1: Motivation and system
466 description. *Wea. Forecasting*, 37(8), 1371–1395. 10.1175/WAF-D-21-0151.1

467 Freitas, S., and Coauthors, 2007: Including the sub-grid scale plume rise of vegetation fires in
468 low resolution atmospheric transport models. *Atmos. Chem. Phys.*, 7(13), 3385–3398.
469 https://doi.org/10.5194/acp-7-3385-2007

Gelaro, R., and Coauthors, 2017: The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2), *J. Climate*, 30, 5419–5454, https://doi.org/10.1175/JCLI-D-16-0758.1.

Hoerl, A. and R. Kennard, 1970: "Ridge Regression: Biased Estimation for Nonorthogonal Problems". *Technometrics*. 12 (1): 55–67. doi:10.2307/1267351. JSTOR 1267351.

Hogan, T., and Coauthors, 2014: The Navy Global Environmental Model. *Oceanography*, 27(3), 116–125. https://doi.org/10.5670/oceanog.2014.73

Hyer, E., J. Reid, and J. Zhang, 2011: An over-land aerosol optical depth data set for data assimilation by filtering, correction, and aggregation of MODIS Collection 5 optical depth retrievals. *Atmospheric Measurement Techniques*, 4(3), 379–408. https://doi.org/10.5194/amt-4-379-2011

Johnston, F., and Coauthors, 2012: Estimated Global Mortality Attributable to Smoke from Landscape Fires. *Environ. Health Perspect*. 2012 May; 120(5): 695–701. doi: 10.1289/ehp.1104422.

Kang, D., R. Mathur, K. Schere, S. Yu, and B. Eder, 2007: New Categorical Metrics for Air Quality Model Evaluation. *J. Appl. Meteor. Climatol*., 46, 549–555. https://doi.org/10.1175/JAM2479.1

Koenker, R., and G. Bassett, 1978: Regression Quantiles. *Econometrica* 46, no. 1, 33–50. https://doi.org/10.2307/1913643.

Li, Y., and Coauthors, 2020: Ensemble PM 2.5 Forecasting During the 2018 Camp Fire Event Using the HYSPLIT Transport and Dispersion Model. *J. Geophys. Res. Atmos.*, 125(15). https://doi.org/10.1029/2020JD032768

Li, Y., and Coauthors, 2023: Impacts of estimated plume rise on PM2.5 exceedance prediction during extreme wildfire events: A comparison of three schemes (Briggs, Freitas, and Sofiev). *Atmos. Chem. Phys*., 23, 3083–3101, https://doi.org/10.5194/acp-23-3083-2023.

Lynch, P., and Coauthors, 2016: An 11-year global gridded aerosol optical thickness reanalysis (v1.0) for atmospheric and climate sciences. *Geosci. Model Dev*., 9(4), 1489–1522. https://doi.org/10.5194/gmd-9-1489-2016

Makkaroon, P., and Coauthors, 2023: Development and Evaluation of a North America Ensemble Wildfire Forecast: Initial Application to the 2020 Western United States "Gigafire". *J. Geophys. Res. Atmos*. 128, e2022JD037298. https://doi.org/10.1029/2022JD037298

Pan, X., and Coauthors, 2020: Six global biomass burning emission datasets: Intercomparison and application in one global aerosol model. *Atmos. Chem. Phys*., 20(2), 969–994. https://doi.org/10.5194/acp-20-969-2020

Randles, C., and Coauthors, 2017: The MERRA-2 Aerosol Reanalysis, 1980 Onward. Part I: System Description and Data Assimilation Evaluation. *J. Climate*, 30(17), 6823–6850. https://doi.org/10.1175/JCLI-D-16-0609.1

Reid, J., and Coauthors, 2009: Global monitoring and forecasting of biomass- burning smoke: Description of and lessons from the Fire Locating and Modeling of Burning Emissions (FLAMBE) program. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens*., 2(3), 144–162. https://doi.org/10.1109/JSTARS.2009.2027443

Sessions, and Coauthors, 2015: Development towards a global operational aerosol consensus: basic climatological characteristics of the International Cooperative for Aerosol Prediction Multi-Model Ensemble (ICAP-MME), *Atmos. Chem. Phys*., 15, 335–362, https://doi.org/10.5194/acp-15-335-2015.

Wooster, M. J., Roberts, G., Perry, G. L. W., and Kaufman, Y. J. (2005). Retrieval of biomass combustion rates and totals from fire radiative power observations: FRP derivation and calibration relationships between biomass consumption and fire radiative energy release. *Journal of Geophysical Researc*h, 110, D24311. https://doi.org/10.1029/2005JD006318

Xian, P., and Coauthors, 2019: Current state of the global operational aerosol multi-model ensemble: An update from the International Cooperative for Aerosol Prediction (ICAP). *Q. J. R. Meteorol. Soc*., 145(S1), 176–209. https://doi.org/10.1002/qj.3497

Ye, X., and Coauthors, 2021: Evaluation and intercomparison of wildfire smoke forecasts from multiple modeling systems for the 2019 Williams Flats fire. *Atmos. Chem. Phys*., 21(18), 14427–14469. https://doi.org/10.5194/acp-21-14427-2021

Zhang, L., and Coauthors, 2022: Development and evaluation of the Aerosol Forecast Member in the National Center for Environment Prediction (NCEP)'s Global

Ensemble Forecast System (GEFS-Aerosols v1), *Geosci. Model Dev.*, 15, 5337–5369, https://doi.org/10.5194/gmd-15-5337-2022.

Zhang, X., S. Kondragunta, A. Da Silva, S. Lu, H. Ding, F. Li, and Y. Zhu, 2019: The blended global biomass burning emissions product from MODIS and VIIRS observations (GBBEPx) version 3.1. https://www.ospo.noaa.gov/Products/land/gbbepx/docs/GBBEPx_ATBD.pdf

Responses to Reviewers' Comments

We thank the reviewers for taking the time to review our manuscript and providing their constructive comments. Please find below our point-to-point responses to these comments (in black)

Reviewer #1: In this manuscript the performance of an unweighted ensemble model is compared with 5 individual models and AirNow observations for surface PM2.5 in the year 2020. Authors apply weighted ensemble approach, where the weights are determined using 4 different techniques, and their performance was compared with the unweighted ensemble model during PM2.5 exceedance events over the test period.

I recommend it for publication subject to the authors addressing the minor comments below.

Overall, the manuscript presents an appreciated contribution to the field of wildfire air quality forecasting. A need for this type of analysis exists; it has been conveyed from many in the community that a simple multi-model mean, where all models are equally weighted, is not sufficient. The approach introduced here is relevant to a wide range of scientific disciplines and AMS readers.
The manuscript is well-structured, concise, and straightforward, providing a clear presentation of the research methods, results, and their implications. However, in my opinion, the introduction could be improved by including more introductory details and being more thorough in it's literature review and using reference to better backup some statement. In addition, some sections of the results are only quantitative using statistical metrics, where I think a better job on being qualitative on it's description and explanation of the results could have been done. More details in the comments below.

First of all, I was wondering if there's a specific reason for not including Bluesky-HYSPLIT in this comparison. Many in the community would find it interesting to assess its performance alongside these models.

Response: Thank you for the comment. In our previous study, we included NOAA Air Resources Lab (ARL)'s HYSPLIT as one member in our ensemble (Makkaroon et al., 2023) forecast. However, the ARL HYSPLIT disconnected the operations forecast, ARL is working on this issue. Therefore, we skip NOAA ARL HYSPLIT in this study. For the Forest Service Bluesky-HYSPLIT forecast, we have reached out to see if they want to join our ensemble forecast project. We will include either Bluesky-HYSPLIT or ARL HYSPLIT in our future study.

Could you provide references for the following:
- Line 72: "Ensemble forecasting does not work best all the time."
- Line 75: "Moreover, if individual models in the ensemble are biased, the ensemble itself may exhibit systematic bias."
- Line 90: "which is highly correlated with fire emissions, across the 10 U.S. Environmental Protection Agency (EPA) regions for 2022"

Response: Thank you for the comment. To address this comment, we have changed lines 71-76 to:

“While multi-model ensemble often outperforms single-model forecasts, some challenges remain. The ensemble mean does not work best all the time (Xian et al., 2019; Makkaroon et al., 2023). For instance, insufficient diversity among models in the multi-model ensemble can limit the ability of the ensemble to capture the full uncertainties and variability tied to different inputs and assumptions. Moreover, if individual models in the ensemble are biased, the ensemble itself may exhibit systematic bias (DelSole et al., 2016).”

Also, line 90-94 have been changed to:
“Figure 1 displays the annual and monthly total fire radiative energy (FRE) from Global Biomass Burning Emissions Product (GBBEPx; Zhang et al., 2019), which is highly correlated with fire emissions (Wooster et al., 2005), across the 10 U.S. Environmental Protection Agency (EPA) regions for 2022.”

Line 148: add the spatial resolution for GEOS-FP model as well.
Response: Thank you for the comment. Line 155-156 have been changed to:
“This study used the GEOS Forward Processing system (GEOS-FP, version 5.27.1) at a 0.25° × 0.3125° spatial resolution…”

Line 236: Can you explain the reasoning behind the specific choice of assigning a weight of 10 vs 1? and can you clarify why a threshold-based approach to assign weights? I know we are seeking for extreme (wildfire) events, but I am wondering if bin of values in assigning weights would not be more appropriate (considering observation and model's uncertainties)?

Response: Thanks for the comments. We conducted tests with varying weights for extreme events, ranging from 2 to 100. Our findings indicate that a weight below 10 did not yield significant improvement in extreme events forecasting. Conversely, weights exceeding 100 led to increased computing time without substantial enhancement in extreme events forecasting. Consequently, we opted for a weight of 10. The 20 $\mu g/m^3$ (top 20 %) threshold is based on Kang et al., 2007, which they used as the basis for calculating the weighted success index in extreme events forecast. The threshold for $PM_{2.5}$ exceedance is 35 $\mu g/m^3$, we use 20 $\mu g/m^3$ to include more cases considering observation and the model’s uncertainties.

To make it clear, lines 251-252 have been changed to:
“…80% of the total observations, based on Kang et al. (2007), the basis for calculating the weighted success index for extreme forecast…”

Line 263: You mentioned that the order of models in the results section differs from their original sequence in section 2b, but you haven't specified which model corresponds to each number here.

Response: Thank you for your comments. We deliberately adjusted the model order in this section, listing models 1-5 instead of providing the specific names of each model. This intentional rearrangement aims to prevent any distraction from the individual model performances, allowing for a more focused evaluation of the ensemble forecast. The purpose of this paper is to assess the performance of the ensemble mean and various weighted ensemble methods, rather than delving into individual model assessments. The evaluation of the individual models can be found in the reference of each model.

Lines 281-284 have been changed to:
"The purpose of this study is to assess ensemble forecast skills rather than delving into the performance of individual models. We intentionally rearranged the order of these models and renamed them to models 1-5 to avoid focusing on specific model performance."

Line 278: When you describe the results in Fig. 3, it would be beneficial to provide a more detailed explanation of the findings discuss your insights into why some models performs differently in various regions compared to the ensemble approach.
Response: Thank you for your comments. The ensemble mean is overperformed by some models because of the underestimation of the anthropogenic emission. We have added the Seasonal PM2.5 forecast average overlaid with AirNow observations to the Supplemental Material.

Line 308-312 have been changed to:
"In regions 6, 7, and 9, the scores are mostly negative with some transient positive scores in early 2022 as well as some positive tendency at the end of the year (because of the underestimate of the anthropogenic emission, Figure S2-5), indicating that MMA performs better than each model most of the time, especially in the wildfire season (Summer and Fall)."

The following figures have been added to the Supplemental Material：

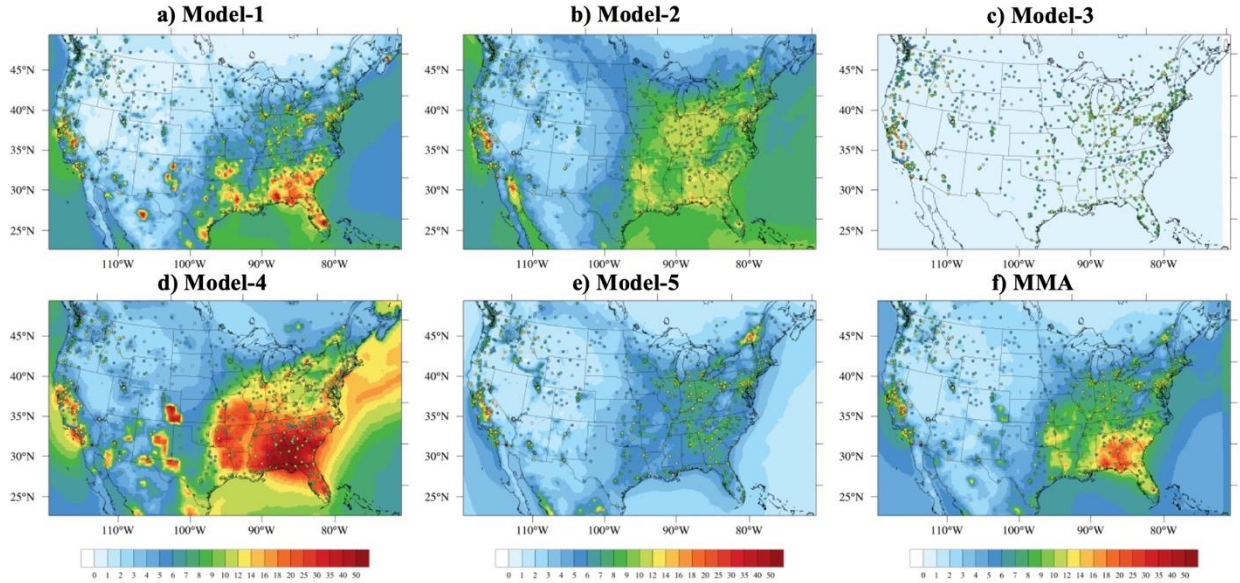

Fig. S2. January to March averaged surface $PM_{2.5}$ concentration (contour) predicted by models 1 to 5 and MMA and observed by the AirNow network (colored circles) for 2022.

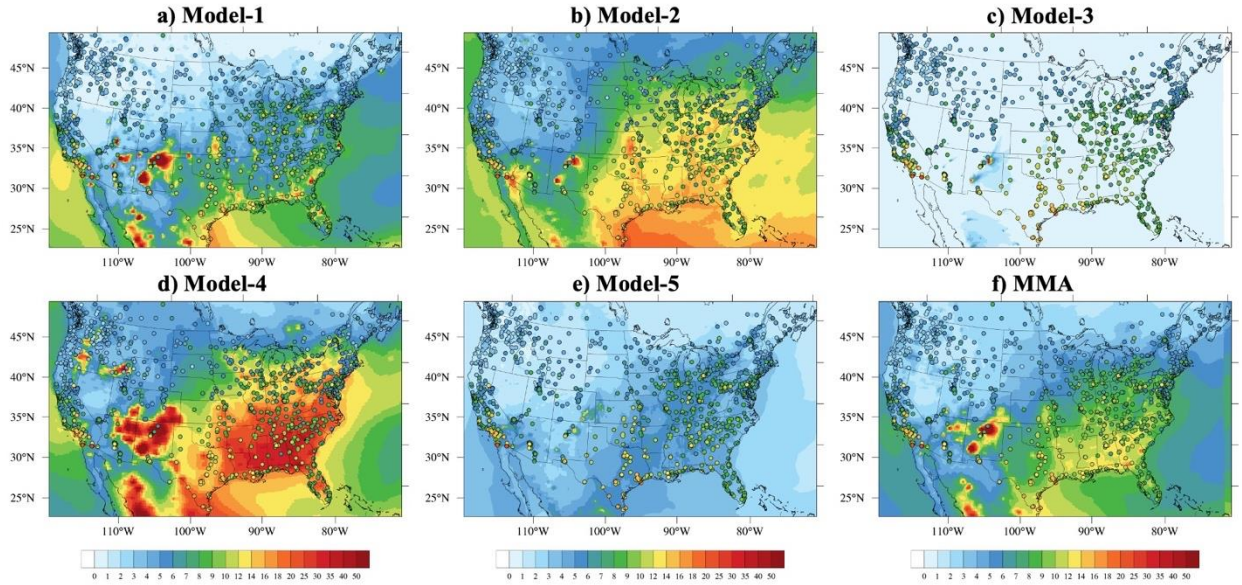

Fig. S3. April to June averaged surface $PM_{2.5}$ concentration (contour) predicted by models 1 to 5 and MMA and observed by the AirNow network (colored circles) for 2022.

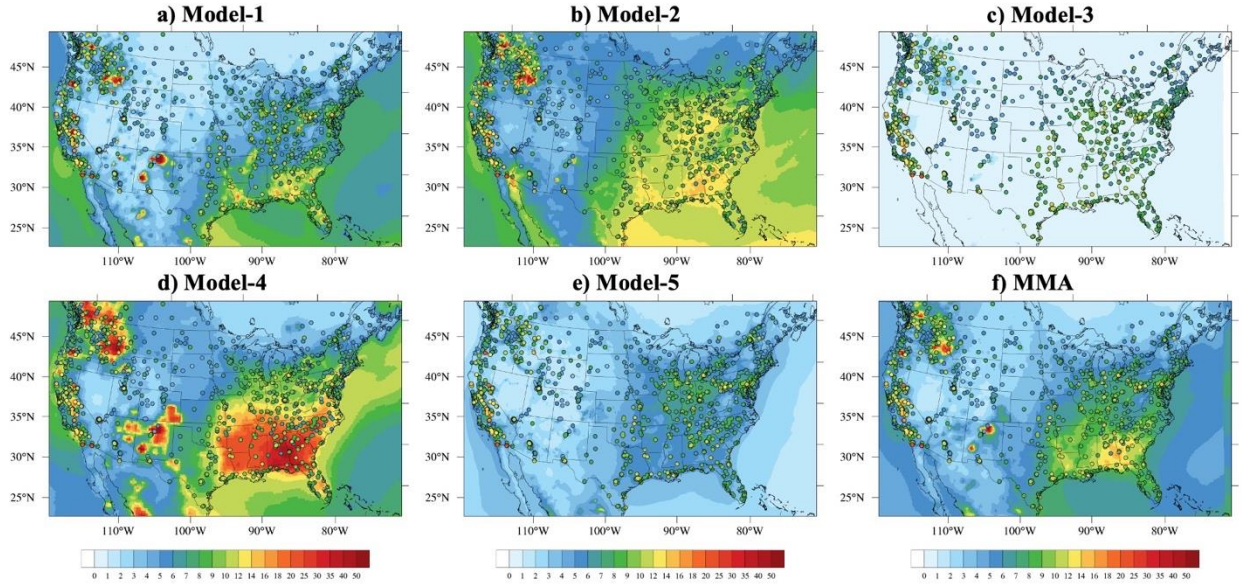

Fig. S4. July to September averaged surface $PM_{2.5}$ concentration (contour) predicted by models 1 to 5 and MMA and observed by the AirNow network (colored circles) for 2022.

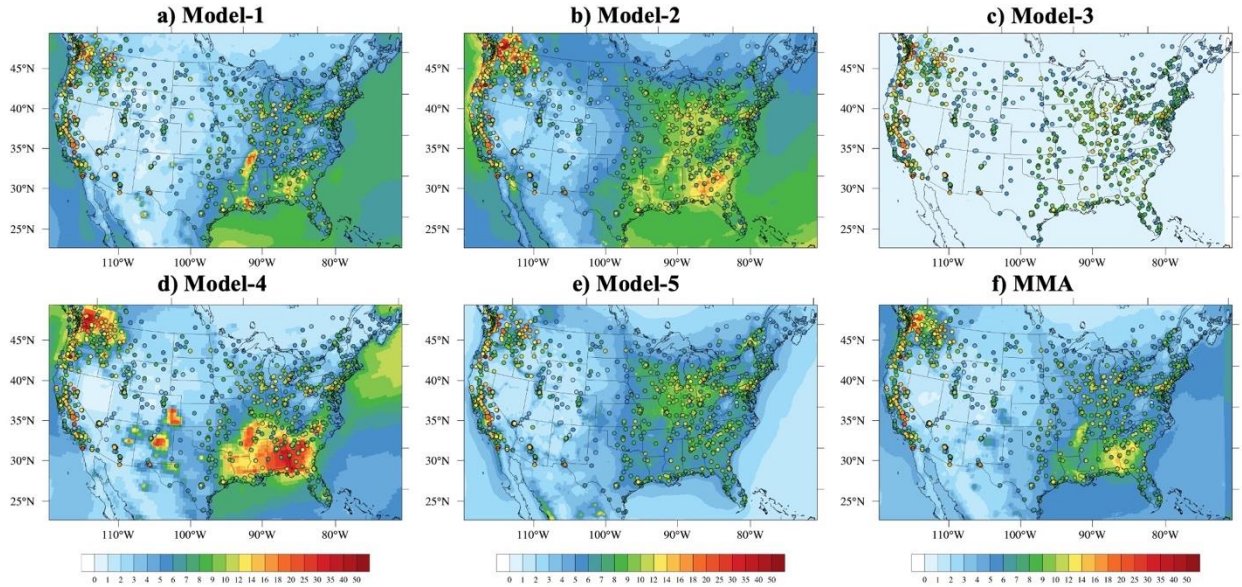

Fig. S5. October to December averaged surface $PM_{2.5}$ concentration (contour) predicted by models 1 to 5 and MMA and observed by the AirNow network (colored circles) for 2022.

Line 311: In the legend for Table 2, when you refer to 'ensemble (MMA) forecasts,' please specify you are referring to weighted and unweighted ensemble MMA forecasts. It appears that you use 'MMA' interchangeably with 'unweighted ensemble' in different parts of the manuscript, such as in line 306 (unweighted ensemble mean, MMA) and line 265 (unweighted MMA ensemble). To avoid confusion, it would be helpful to maintain consistency in your usage of this abbreviation throughout the manuscript.
Additionally, you introduce abbreviations at a few places in the manuscript, including lines 180 and 263. Similarly, for MLR mentioned in lines 195, 197, 306, and 348. There is also no need to reintroduce these abbreviations; WR in line 355 and RR in line 530.
Response: Thanks for the comments. MMA means Multi-Model Average. In the whole paper, it refers to the ensemble mean. We agree that the Table 2 caption is confusing. It has been changed to:

"Table 2. The Fractional bias (FB), aH, aFAR, and WSI for the different models (Model-1 to Model-5, M1-M5), MMA, and four weighted ensemble forecasts (MLR, RR, QR, and WR) for the October to December 2022 testing period (bold represents the best results, underline represents the worst results)."

We have modified several places in the paper to address the abbreviation concerns.

Line 410: The variables in equation A5 are not introduced.
Response: Thanks for the comments. The following sentence has been added to Lines 479-480: "where a, b, and c refer to the three scenarios defined above, and n represents the total number of observations."

Reviewer #2: Wildfires are a major concern for the public, but because major events occasionally erupt, it's difficult for models to accurately predict them in a timely manner. The ensemble approach discussed in this article is a promising approach. The article is well organized and of practical scientific interest. It is therefore worthy of acceptance and publication.
I have two main questions that I would like the author to clarify before publication.
1. The five models author selected for this study vary widely and include global and regional models. I would expect their simulation results to differ significantly as well. Figure 2 is only the annual average surface PM2.5 concentration and is not enough to truly show their differences. Can the author add more discussion to this? Because I want to know if the ensemble approach talked about in this manuscript only applies to this case. For real-time wildfire forecasting, it is almost impossible to collect model forecast data from different agencies at the same time. It is typically one model with multiple ensemble members driven by model perturbations. The differences in model simulations should be much smaller than the cases discussed in this study.

Response: Thanks for the comments. We add more results in the Supplemental Material. The purpose of this paper is to assess the performance of the ensemble mean and various weighted ensemble methods, rather than delving into individual model assessments. Therefore, we did not investigate the performance of each model. The evaluation of the individual models can be found in the reference of each model.

The ensemble forecast is running in real-time, not only for the case demonstrated in this paper. We set up automatic transfer links between our server and the server of each model. The ensemble forecast is conducted automatically after we get all data from each individual model. We have added more explanation in the paper.

The following figures have been added to the Supplemental Material：

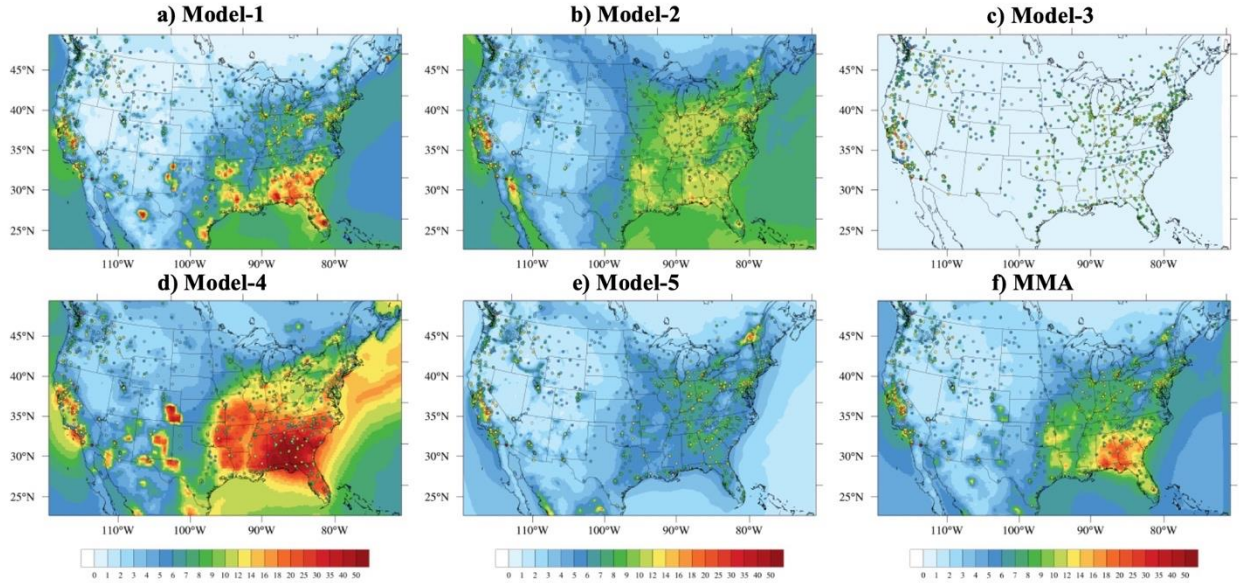

Fig. S2. January to March averaged surface $PM_{2.5}$ concentration (contour) predicted by models 1 to 5 and MMA and observed by the AirNow network (colored circles) for 2022.

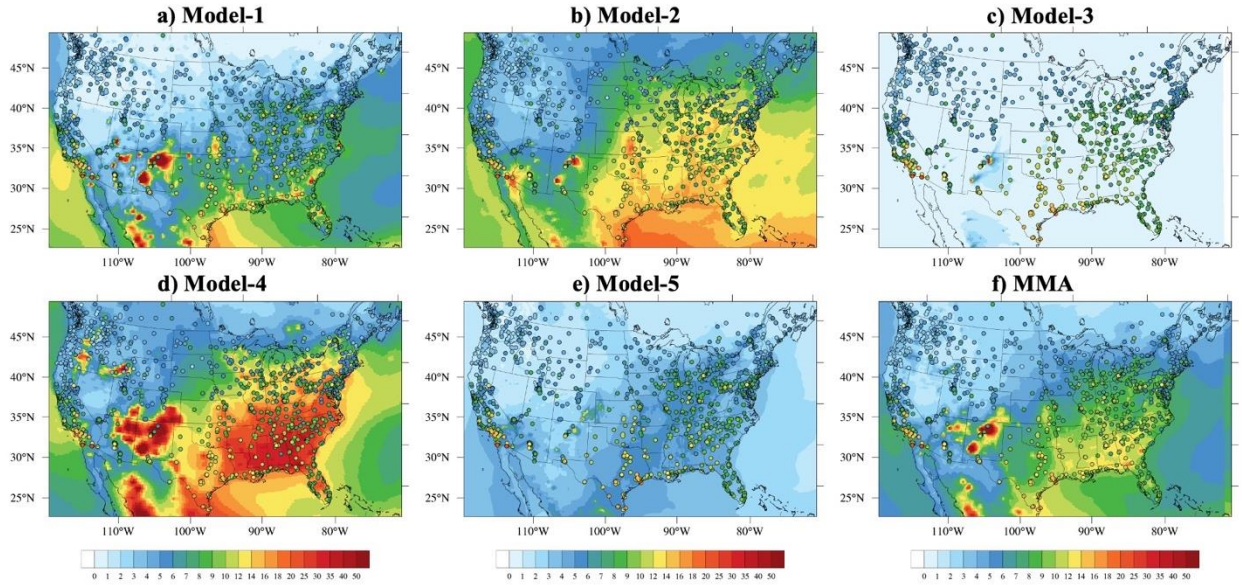

Fig. S3. April to June averaged surface $PM_{2.5}$ concentration (contour) predicted by models 1 to 5 and MMA and observed by the AirNow network (colored circles) for 2022.

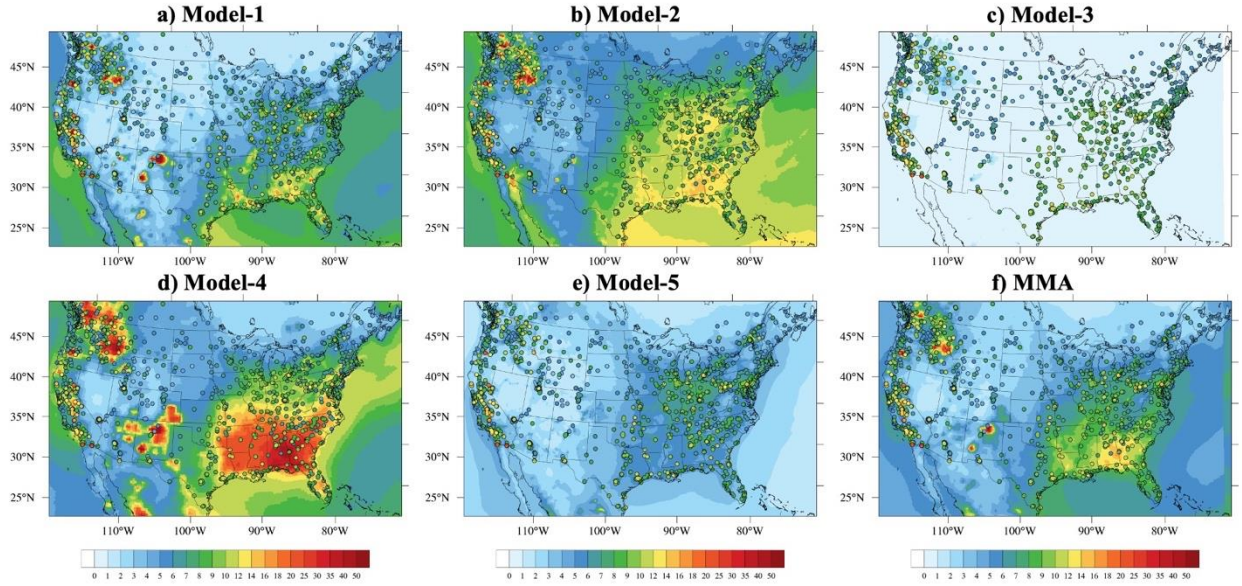

Fig. S4. July to September averaged surface $PM_{2.5}$ concentration (contour) predicted by models 1 to 5 and MMA and observed by the AirNow network (colored circles) for 2022.

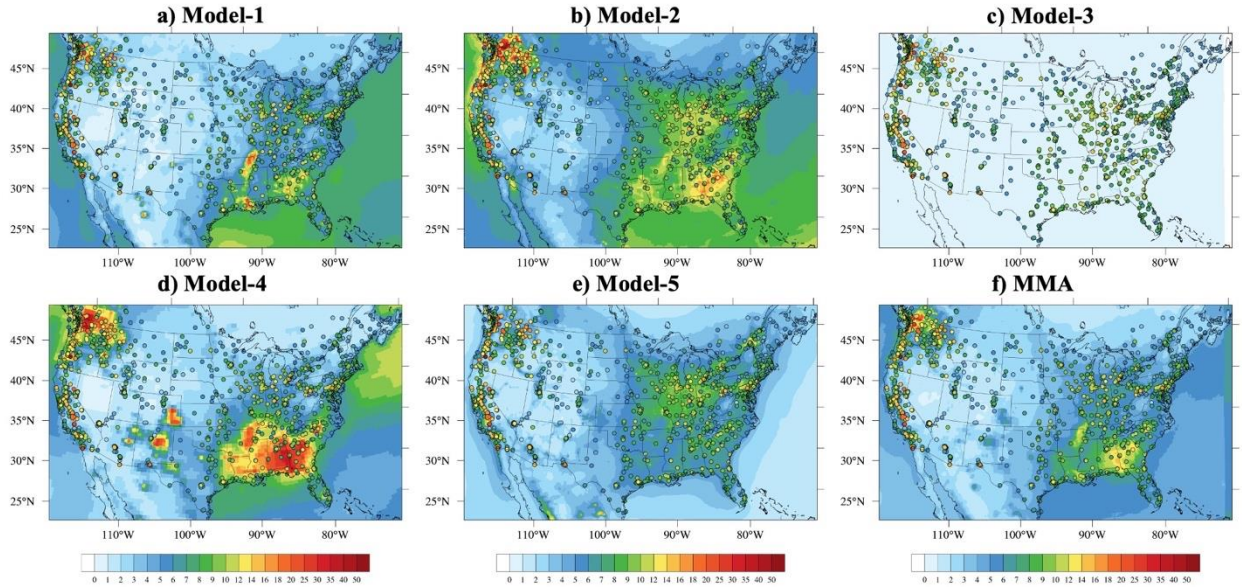

Fig. S5. October to December averaged surface $PM_{2.5}$ concentration (contour) predicted by models 1 to 5 and MMA and observed by the AirNow network (colored circles) for 2022.

The following sentence has been added to Line 401-403:
"Automated transfer links have been established between these agencies and GMU. Individual model daily forecast results are automatically transmitted to GMU each day to generate the real-time ensemble forecast results."

2. I would like to know how the weights (β) and ridge parameters (λ) were calculated in this study. Because the author used one year of data to test the method, with the first 9 months of data as training data and the last 3 months of data as test data. However, by comparing the model evaluation matrix shown in Table 1 and Table 2, I noticed that the statistics for the same model are very different. Is this caused by using observation data from different time periods? If the answer is yes, then a key question is can we use the current weights and ridge parameters in future situations?

Response: Thanks for the comments. The weights are calculated using the Matlab ridge regression function "ridge" (https://www.mathworks.com/help/stats/ridge.html). We tested the value of λ from 1 to 1000. We used the weights and ridge parameters calculated from the training period to calculate the weighted ensemble for the testing period.

The results in Table 1 are based on the forecast for the whole year, and the results in Table 2 are based on the testing period (October to December). Although the values of aH, aFAR, and WSI look very different between the two tables for the same model, the patterns are similar. The models with high aH (low aFAR, high WSI) values in Table 1, also have the high aH (low aFAR, high WSI) values in Table 2. It is okay for the absolute values to differ across the periods as long as the relative skills remain similar.

The following sentence is added to line 225:
"We tested the value of λ from 1 to 1000."

Here are some minor issues

* Please replace "Model-x" with the model's real name

Thank you for your comments. We deliberately adjusted the model order in this section, listing models 1-5 instead of providing the specific names of each model. This intentional rearrangement aims to prevent any distraction from the individual model performances, allowing for a more focused evaluation of the ensemble forecast. The purpose of this paper is to assess the performance of the ensemble mean and various weighted ensemble methods, rather than delving into individual model assessments. The evaluation of the individual models can be found in the reference of each model.

Lines 281-284 have been changed to:
"The purpose of this study is to assess ensemble forecast skills rather than delving into the performance of individual models. We intentionally rearranged the order of these models and renamed them to models 1-5 to avoid focusing on specific model performance."

* For Figures 3 and 4, please briefly explain why only EPA Regions 4, 6, 7, 9, and 10 are shown.

Response: Based on Figure 1c, the EPA Regions 4, 6, 7, 9, and 10 are the major fire regions in the year 2022. Since this paper focuses on the wildfire air quality ensemble forecast, we show the results of regions 4, 6, 7, 9, and 10 in Figs 3 and 4.

Lines 301-302 have been changed to:
"We compared the MMA with each individual model using the random walk method for major fire regions (EPA region 4, 6, 7, 9, and 10 from Fig. 1c; Fig. 3)"

* Line 183, please define β0

Response: The following sentence has been added to Line 193:
"… and β_0 is the intercept."

* Lines 209-210, where is equation A10?

Response: Sorry, it is a typo. It should be equation (3). It has been changed to Eq3 in the revision.

* Lines 282-283, I don't see a positive score for area 9 in Figure 3

Response: Thanks for the comments. For the random work method, positive tendency means the individual models work better in that period. There are positive tendencies at the end of 2022 for Region 9. The positive value illustrates the cumulative performance (overall averaged performance) before that period.

To make it clear lines 302-306 have been changed to:
"Figure 3 shows the relative forecast score of individual models compared to MMA. A negative value on day n indicates that the overall performance of MMA surpasses that of the individual model from day 1 to day n; the negative trend observed from day n_1 to n_2 signifies that the MMA consistently outperforms the individual model between n_1 and n_2, and vice versa."

* In Tables 1 and 2, some numbers are underlined. What does this mean?

Response: Thanks for the comments. In Tables 1 and 2, the best results are highlighted in bold, while the worst results are underlined. This is mentioned in the Table 1 caption but forgotten in the Table 2 caption. The caption of Table 2 has been changed to:
"Table 2. The Fractional bias (FB), aH, aFAR, and WSI for the different models (Model-1 to Model-5, M1-M5), MMA, and four weighted ensemble forecasts (MLR, RR, QR, and WR) for the October to December 2022 testing period (bold represents the best results, underline represents the worst results)."

* In Table 2, can the authors explain why QR scores higher than WR?

Response: Thank you for your feedback. In this paper, we employ QR and WR methods to enhance the weighted ensemble forecast, particularly for extreme events. QR focuses on the top 10% of cases, whereas WR assigns greater weight to the top 20% of cases. Consequently, predictions using QR tend to be higher than WR. QR exhibits more overestimation and less underestimation compared to WR. As reflected in Table 2, this is why the Fractional Bias of QR is higher (15%) than WR, the area Hit rate of QR is higher (17%), and the false alarm rate is also elevated (55%) compared to WR. We also added a figure to the Supplemental Material to compare QR with WR.

The following lines have been added to the paper line 383:
"QR focuses on the top 10% of cases, whereas WR assigns greater weight to the top 20% of cases. Consequently, predictions using QR tend to be higher than WR (Fig. S6). QR exhibits more overestimation and less underestimation compared to WR. As reflected in Table 2, the Fractional Bias of QR is 15% higher than WR, the area Hit rate of QR is 17 % higher, and the false alarm rate is also elevated (55%) compared to WR. WR offers a balanced enhancement."

The following figure has been added to the Supplemental Material.

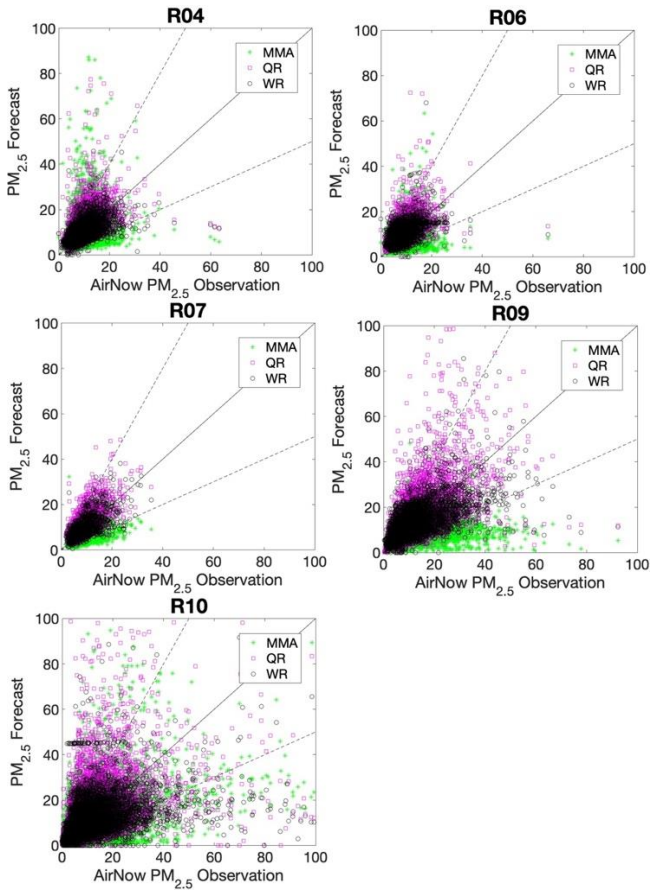

Fig. S6. Scatter plots between predicted and observed $PM_{2.5}$ for MMA (green), QR (magenta), and WR (black) for five fire-prone EPA regions. The solid black line represents the 1:1 ratio line for the observations and forecasts, while the dashed black lines represent the 1:2 and 2:1 ratio lines.

Reviewer #3: Review Comments for "Development of a Real-time Hazardous Air Quality Ensemble System for Improved Wildfire Air Quality Forecasting" submitted by Li et al.
General Comments:

Wildfire events present a significant challenge to air quality (AQ) predictions due to model system incompleteness and uncertainties in fire emission treatment. The manuscript presents the development of the Hazardous Air Quality Ensemble System (HAQES) and its success in improving PM2.5 predictions during wildfire events over the CONUS. The approach of using a weighted ensemble forecast contributes to improved forecasting accuracy, along with the application of quantile regression and weighted regression methods. The manuscript effectively compares the HAQES ensemble forecast with individual models, showcasing a substantial enhancement in forecast accuracy. The inclusion of quantitative measures, such as fractional bias, false alarm rate, and hit rate, strengthens the study's credibility. Overall, the manuscript is well-structured, providing a clear overview of the research, including the problem statement, methodology, and key findings. However, several concerns need
further clarification, and the manuscript is recommended for publication in BAMS upon addressing these comments.

Specific Comments:
1. While the manuscript successfully addresses the impact of heavy wildfire events on air quality predictions, the use of 2022 as an example, being a fire-inactive year, raises questions. It might be beneficial to consider other fire-active years to better demonstrate the HAQES's advantages over individual models during intense wildfire events.
Response: Thanks for the comments. The HAQES will be real-time in the future. Therefore, we only use real-time forecast models to create the ensemble. However, some models were not operational or had missing periods before 2021. So, we use the year 2022 as an example. When we wrote the paper, the data for the year 2023 were not ready. Also, we have limited space to save the data. That's why we only use one year's data in this paper. The fire activity of 2022 is around the average of 2001-2022 based on the NOAA NCEI report (https://www.ncei.noaa.gov/access/monitoring/monthly-report/fire/202213). The acres burned is the 11th most of the 23 years, the number of fires is the 12th most, and the acres burned per fire is the 10th most. In the future, we will test the performance of HAQES for more fire-active years.

Line 89-90 have been changed to:
"This paper focuses on the year 2022 when 66,255 fires (12th most since 2001) burned 7,534,403 acres (11th most), as reported by the National Interagency Fire Center."

2. In the abstract, the authors assert that weighted ensemble reduced fractional bias by 34%, false alarm rate by 72%, and increased hit rate by 17%. However, in Section 3.2, it is noted that MLR reduces the fractional bias by 34%, increases the hit rate by 17%, and reduces the false alarm rate by 72%. It is recommended to explicitly specify the Multilinear Regression (MLR) in the abstract to avoid confusion since MLR is one of the weighted ensembles.

Response: Thanks for the comments. We have modified the paper accordingly. We have added the "multilinear regression" to the abstract. Lines 27-28 have been changed to:

“Compared to the unweighted ensemble mean, the multilinear regression weighted ensemble reduced fractional bias by 34% …”

3. In the text, it is stated that the first 9-month simulations serve as the training data, with the subsequent three months designated for testing (see Lines 187-188). However, there appears to be a discrepancy, as Table 2 displays results for October-December, while Figures 1-3 showcase outcomes for the entire year of 2022. To avoid potential confusion for readers, please provide clarification on this inconsistency.

Response: Thanks for the comments. The 9-month training data and 3-month testing data are only used for the weighted ensemble since we need some training data to calculate the weight of each model. The MMA does not need any training. Therefore, for Fig. 1-3 and Table 1, we show the results for the whole year, and for Table 2 and Fig. 4, we show the results for the testing period.

The following lines have been added to lines 338-340:
“As explained in section 2.d, the initial 9 months are utilized for weight calculation, while the subsequent 3 months serve as the testing data, which will be assessed in this section.”

4. Should the results for quantile regression (QR) and weighted regression (WR) be incorporated into Figure 4 for a comprehensive visual representation, or would it be more suitable to present them in a separate figure in the appendix to ensure clarity and enhance the visualization of these specific QR and WR findings?

Response: Thanks for the comments. We have modified the paper accordingly. We have added more discussion about QR and WR in section 3. b, and we have added 1 more figure comparing WR and QR to the Supplemental Material.

The following lines have been added to the paper line 383:
“QR focuses on the top 10% of cases, whereas WR assigns greater weight to the top 20% of cases. Consequently, predictions using QR tend to be higher than WR (Fig. S6). QR exhibits more overestimation and less underestimation compared to WR. As reflected in Table 2, the Fractional Bias of QR is 15% higher than WR, the area Hit rate of QR is 17 % higher, and the false alarm rate is also elevated (55%) compared to WR. WR offers a balanced enhancement.”

The following figure has been added to the Supplemental Material.

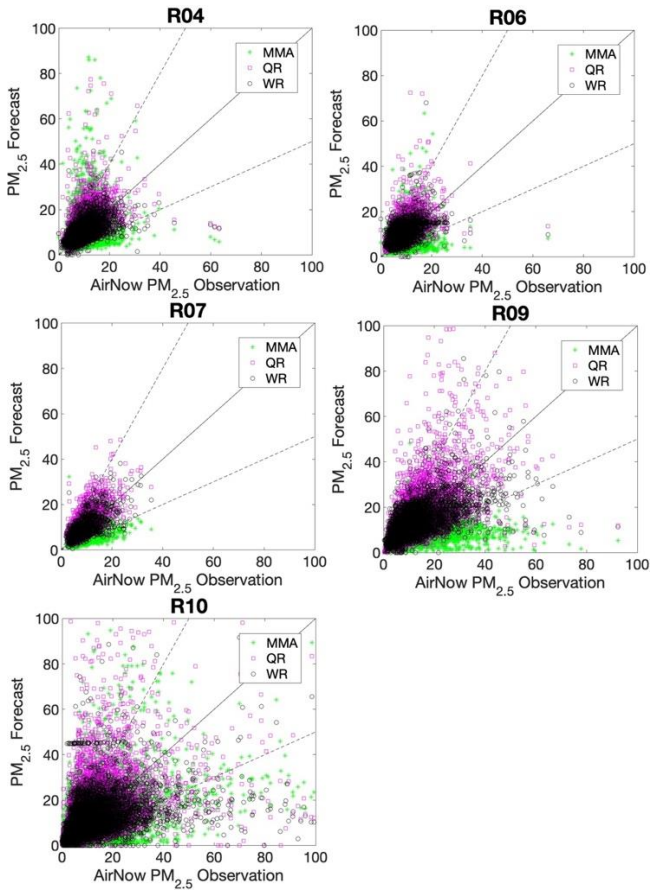

Fig. S6. Scatter plots between predicted and observed $PM_{2.5}$ for MMA (green), QR (magenta), and WR (black) for five fire-prone EPA regions. The solid black line represents the 1:1 ratio line for the observations and forecasts, while the dashed black lines represent the 1:2 and 2:1 ratio lines.

5. In the abstract (Line 23), the authors assert the introduction of a new real-time Hazardous Air Quality Ensemble System (HAQES) in the manuscript. However, it appears that the presented results pertain to the year 2022 rather than real-time outcomes. Please clarify how the five operational forecasts are integrated into the HAQES system to generates ensemble forecast products in real-time.

Response: Thanks for your comments. We have modified the paper accordingly. The following sentence has been added to Line 401:
"Automated transfer links have been established between these agencies and GMU. Individual model daily forecast results are automatically transmitted to GMU each day to generate the real-time ensemble forecast results."

6. Are you sure that the WRF-Chem's atmospheric aerosol chemistry is used in the GEFS-Aerosol model (Line 133)?

Response: Based on the information from GEFS-Aerosol website (https://www.arl.noaa.gov/research/surface-atmosphere-exchange-home/tools-and-products/gefs-

aerosols/#:~:text=Global%20Ensemble%20Forecast%20System%20(GEFS)%20–%20Aerosol%20Model&text=The%20aerosol%20component%20of%20atmospheric,and%20Transport%20model%20(GOCART).) the aerosol component of atmospheric composition in the GEFS is based on the Weather Research and Forecasting model coupled with Chemistry (WRF-Chem).

-Barry, could you answer this question?

7. Consider using the number of Airnow observation sites rather than the number of cities for more meaningful representation (Line 176).

Response: Thanks for the comments. We have modified the paper accordingly.
Lines 183-184 have been changed to:
"It contains air quality data for more than 500 cities across the U.S. (total of 1156 sites), as well as for Canada and Mexico."

8. Please make sure abbreviations are consistently and appropriately defined. For example, GOCART and MMA are both defined twice, while NPP and MODIS are never defined.

Response: Thanks for the comments. We have modified several places in the paper to address the abbreviation concerns.

9. Provide clarification on why β□ is included in Eq. (1) and what value of β□ is used in this study (Line 183).

Response: Thanks for the comments. β_0 is the intercept calculated by MATLAB functions and exhibits variations across the ~1000 AirNow sites. As a result, we have chosen not to list all the specific β_0 values in the paper.

The following sentence has been added to Line 193:
"… and β_0 is the intercept."

10. Explain the choice of 20 µg/m3 in Eq. (7) instead of the EPA standards for PM2.5 (35 µg/m3 for the 24-hour average and 12 µg/m3 for the annual average).

Response: Thanks for the comments. The 20 µg/m^3 (top 20 %) threshold is based on Kang et al., 2007, which they used as the basis for calculating the weighted success index in extreme events forecast. The threshold for $PM_{2.5}$ exceedance is 35 µg/m^3, we use 20 µg/m^3 to include more cases considering observation and the model's uncertainties. To make it clear, lines 251-252 have been changed to:
"…80% of the total observations, based on Kang et al. (2007), the basis for calculating the weighted success index for extreme forecast…"

11. Please correct the inconsistencies in units in Lines 233 and 236 to be µg/m3.

Response: Line 250 has been changed to "…higher than 20 µg/m^3…".

12. Please keep consistency in notation, using Kα in Lines 246-247 to match Eq. (A2).

Response: Line 264 has been changed to "…A significance test (K_α, Appendix B) is conducted to show if A is significantly better ($K > K_\alpha$) or worse ($K < N - K_\alpha$) than B.…".

13. Please specify the parameter presented on the Y-axis in Figure 3's caption and clarify why negative values indicate that the ensemble forecast is more skillful (Lines 279-280).

Response: Thanks for the comments. The Y-axis title has been added to Fig.3. The following lines have been added to Line 302 to clarify why negative values indicate that the ensemble forecast is more skillful:
"Figure 3 shows the relative forecast score of individual models compared to MMA. A negative value on day n indicates that the overall performance of MMA surpasses that of the individual model from day 1 to day n; the negative trend observed from day n_1 to n_2 signifies that the MMA consistently outperforms the individual model between n_1 and n_2, and vice versa"

14. Line 335: Please write PM2.5 in a subscript way to be consistent with other places.
Response: Thanks for the comments. We have modified it accordingly.