BAMS Article

Multiagency 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] and Peng Xian[f]

KEYWORDS: Ensembles; Forecasting; Aerosols/particulates; Air quality; Biomass burning; Wildfires

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 have 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.

DOI: 10.1175/BAMS-D-23-0208.1
Corresponding authors: Yunyao Li, yli74@gmu.edu; Daniel Tong, qtong@gmu.edu
Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/BAMS-D-23-0208.s1.

Manuscript received 8 August 2023, in final form 15 March 2024, accepted 22 March 2024

AFFILIATIONS: [a] Department of Atmospheric, Oceanic and Earth Sciences, George Mason University, Fairfax, Virginia; [b] Center for Spatial Information Science and Systems, George Mason University, Fairfax, Virginia; [c] NOAA/Air Resources Laboratory, College Park, Maryland; [d] NASA Goddard Space Flight Center, Greenbelt, Maryland; [e] NOAA/Global Systems Laboratory, Boulder, Colorado; [f] Marine Meteorology Division, Naval Research Laboratory, Monterey, California

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) multimodel ensemble (MME), a global operational aerosol multimodel 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 multimodel ensemble forecast system that effectively simulated the 2020 western U.S. "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 multimodel 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 multimodel 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 and Tippett 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.

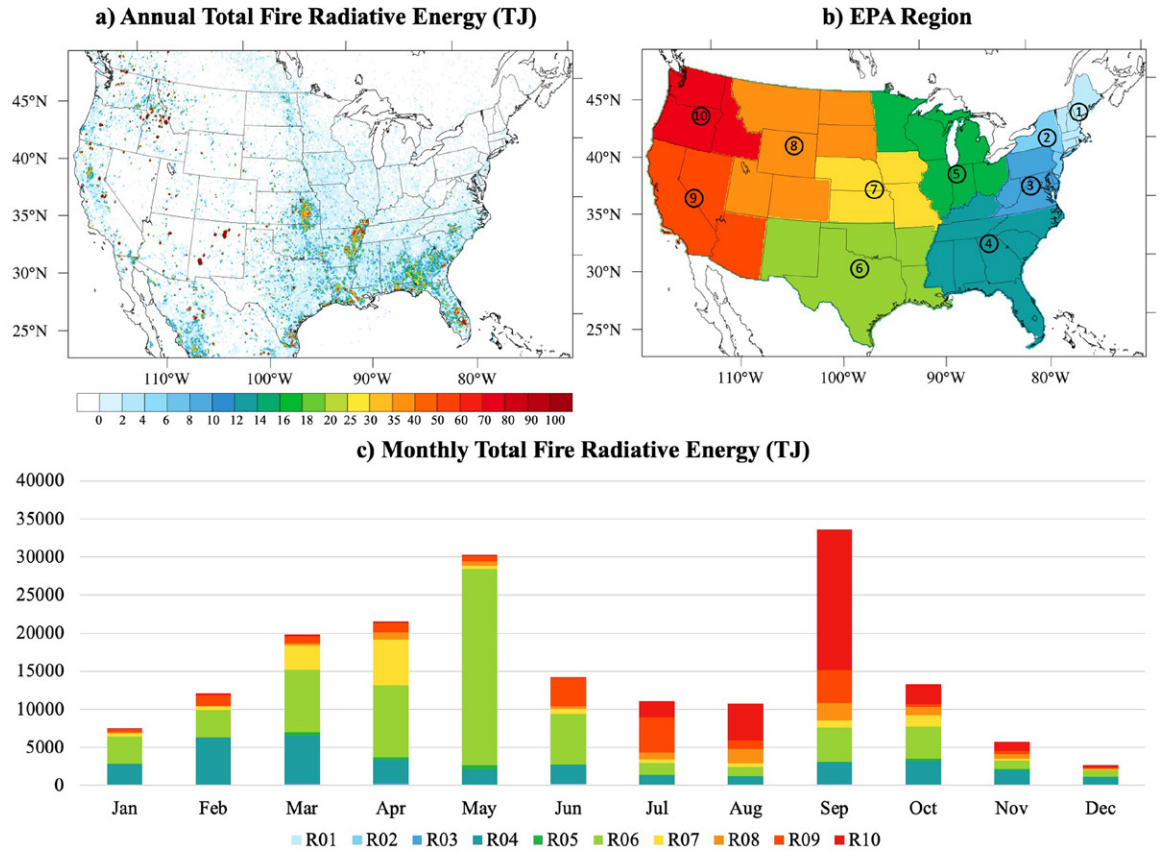

FIG. 1. The (a) annual and (c) monthly total FRE across the (b) 10 U.S. EPA regions for 2022.

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 United States, 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 6281 TJ of fire energy in 1 month. In the central United States, 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 2 months. In the western United States, 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 1 month. Overall, the strongest fire energy occurred in September (33 631 TJ), followed by May (30 330 TJ).

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 global. 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-h surface $PM_{2.5}$

forecasts initialized at 1200 UTC (forecast hour: 0000–2300 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 contiguous 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 (FRP) data from the *Suomi National Polar-Orbiting Partnership* (*SNPP*), *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 nonfire 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 Dynamical Core (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-dimensional 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 the Community Multiscale Air Quality (CMAQ) modeling system 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 (GEOS-FP, version 5.27.1) system at a 0.25° × 0.3125° spatial resolution, which generates analyses, assimilation products, and 10-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 Emission 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 air quality data for more than 500 cities across the United States (total of 1156 sites), as well as for Canada and Mexico.

d. Ensemble design. In this study, we examined five techniques for creating ensembles categorized into two groups: unweighted and weighted ensemble approaches. Unweighted ensemble employed multimodel average (MMA) to merge predictions from multiple models into one consolidated forecast, while weighted ensemble assigned different weights β to member models M_j:

$$\hat{M} = \sum_{j=1}^{S} \beta_j M_j + \beta_0, \tag{1}$$

where S represents the total number of models which is 5 in this study and β_0 is the intercept. To determine the weights, the data for the year 2022 are grouped into training and testing sets. Since wildfires can last for weeks, to ensure the independence of the training and testing data, we did not select the training data randomly. 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 1 year of data, which may lead to variability in the calculated weights for each model. However, 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, to thoroughly investigate its performance and determine the optimal weights for each model.

We experimented with four regression methods to determine these weights: multilinear regression (MLR), ridge regression (RR), quantile regression (QR), and weighted regression (WR).

1) MLR. MLR calculates the weights for each model by minimizing the error between the observation O and the weighted multimodel prediction:

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

where N is the total number of observations.

2) 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:

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

where λ is the ridge parameter. The first 20 days in each month are used to train the data using [(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 multimodel forecasting. It has been found to be useful in climate ensemble studies (DelSole 2007).

3) 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:

$$\hat{M}_{\mathrm{QR}} = \sum_{j=1}^{S}\beta_{j,q}M_j + \beta_{0,q}, \tag{4}$$

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:

$$\hat{\beta}^{\mathrm{QR}} = \arg\min_{\beta}\left[\sum_{j:M\geq M_q} q\left|O_i - \beta_{0,q} - \sum_{j=1}^{S}\beta_{j,q}M_{i,j}\right| + \sum_{j:M<M_q}(1-q)\left|O_i - \beta_{0,q} - \sum_{j=1}^{S}\beta_{j,q}M_{i,j}\right|\right], \tag{5}$$

4) 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:

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

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:

$$W_i = \begin{cases} 10, & \text{if } O_i > 20\ \mu\mathrm{g\,m^{-2}} \\ 1, & \text{otherwise} \end{cases}. \tag{7}$$

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 as follows:

$$d_N = K - \left(N - K\right) = 2K - N. \tag{8}$$

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 U.S. EPA National Ambient Air Quality Standards (NAAQS) 24-h $PM_{2.5}$ exceedance events (>35 μg 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 1–5 (M1–M5); 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.]. Then, 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 1–5 and the MMA is 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.

We compared the MMA with each individual model using the random walk method for major fire regions (EPA regions 4, 6, 7, 9, and 10 from Fig. 1c to 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

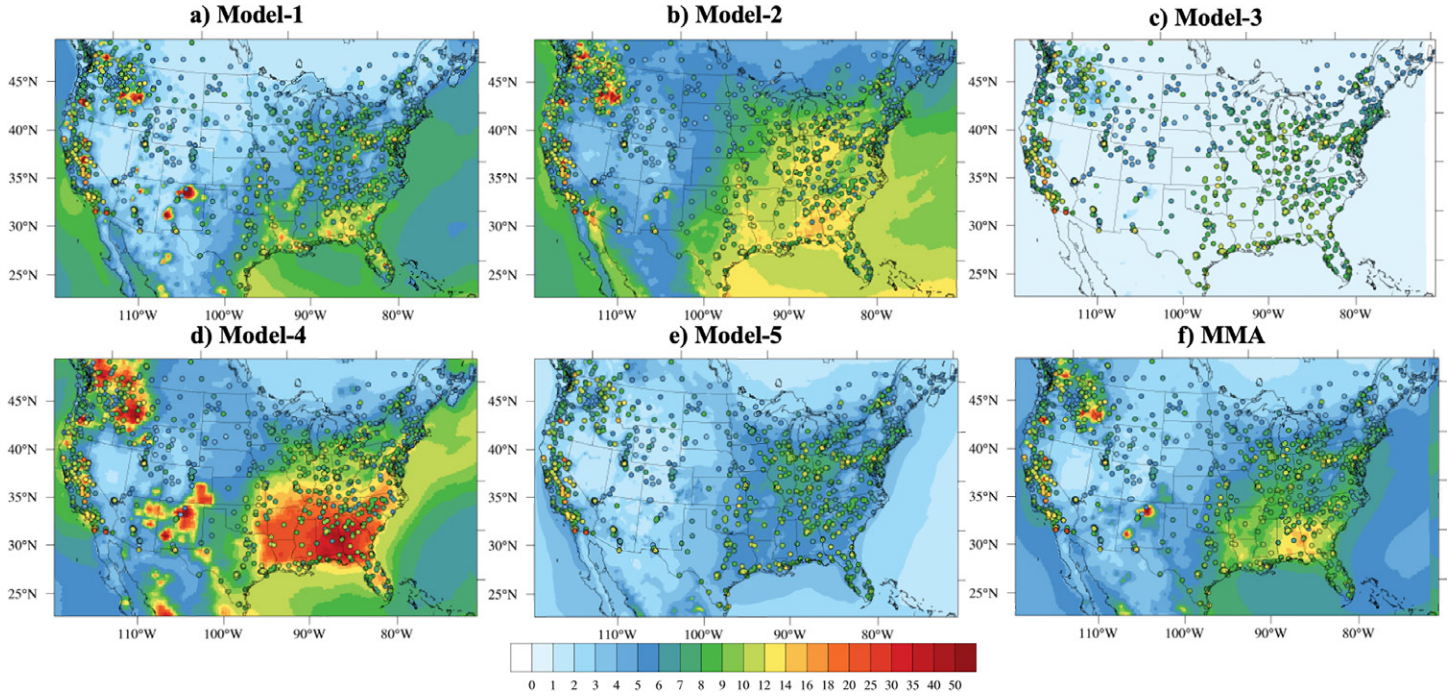

FIG. 2. Annual mean surface $PM_{2.5}$ concentration (contour) predicted by models 1–5 and MMA and observed by the AirNow network (colored circles) for the year 2022.

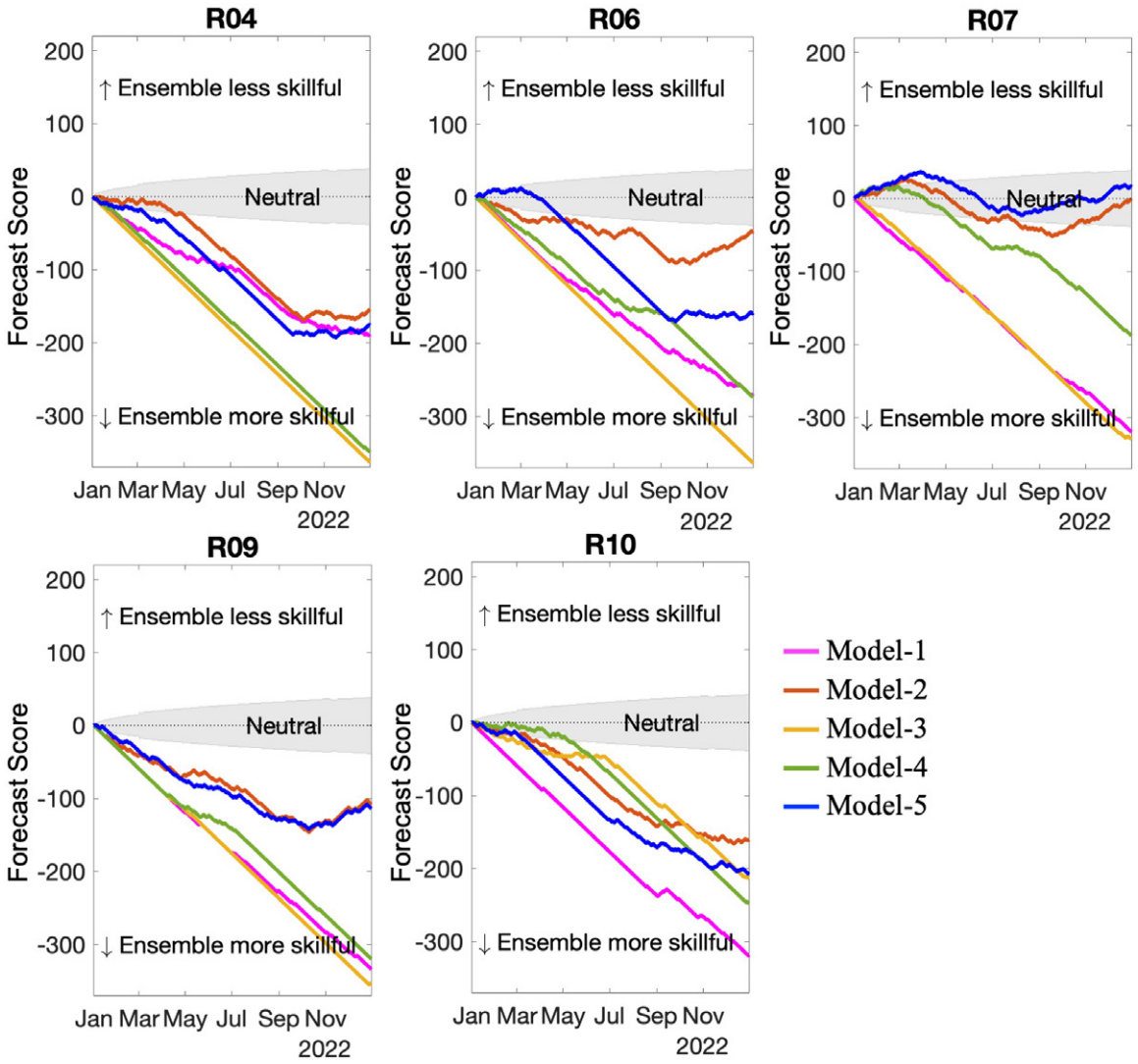

FIG. 3. Comparison of individual models to MMA using the random walk method for major fire regions.

TABLE 1. The aH, aFAR, and WSI for models 1–5 and the MMA for the year 2022. The best results are highlighted in bold, while the worst results are underlined.

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

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 in the online supplemental material). Overall, the MMA outperforms individual models, demonstrating that ensemble forecasts can effectively reduce forecast uncertainty.

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 not only 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).

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 2d, 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%, 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.

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 does not notably enhance model performance. This could be attributed to the modest number of models, so the data are 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 are noisy (DelSole 2007; Pena and van den Dool 2008).

MMA, MLR, and RR all tend to underestimate the $PM_{2.5}$ exceedance events (Fig. 4), particularly on the West Coast with high wildfire emissions (R9 and R10). Therefore, we applied

TABLE 2. The FB, aH, aFAR, and WSI for the different models (M1–M5), MMA, and four weighted ensemble forecasts (MLR, RR, QR, and WR) for the October–December 2022 testing period (bold represents the best results and underline represents the worst results).

	M1	M2	M3	M4	M5	MMA	MLR	RR	QR	WR
FB	0.60	0.49	1.87	0.88	0.50	0.50	**0.33**	0.34	0.41	0.35
aH	42.09	76.87	8.52	65.22	61.04	56.00	65.39	61.04	**86.96**	69.91
aFAR	57.24	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

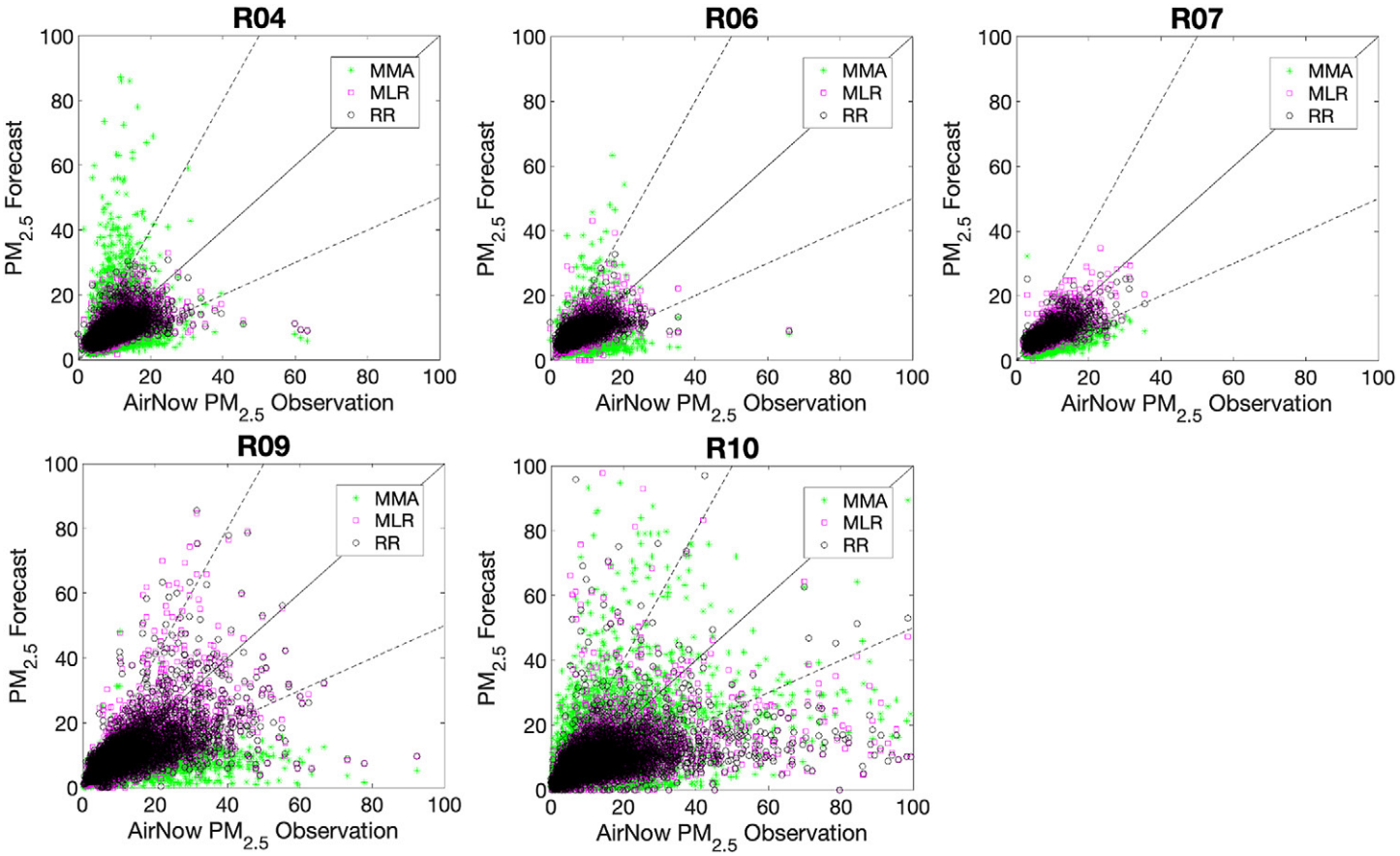

Fig. 4. Scatterplots between predicted and observed $PM_{2.5}$ for MMA (green), MLR (magenta), and RR (black) for five fire-prone EPA regions. The solid black line represents the 1:1 ratio line for the observations and forecasts, while the dashed black lines represent the 1:2 and 2:1 ratio lines.

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 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 George Mason University (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%, reduced the 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. We appreciate the discussions with Dr. Shunjie Tu on ridge regression, quantile regression, and weighted regression. 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 links 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; and NAAPS: https://usgodae.org/pub/outgoing/fnmoc/models; HAQES: http://air.csiss.gmu.edu/haqes. AirNow data can be downloaded from https://files.airnowtech.org/?prefix=airnow/2022/.

APPENDIX A
Fractional Bias

Below is the definition of fractional bias:

$$\mathrm{FB}_i = 2 \times \frac{\left| O_i - M_i \right|}{O_i + M_i}, \tag{A1}$$

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

APPENDIX B
Significance Test K_α for Random Walk

The K_α can be approximated as follows:

$$K_\alpha = \left\lceil \frac{N}{2} - z_{\alpha/2} \sqrt{\frac{N}{4} - \frac{1}{2}} \right\rceil, \tag{B1}$$

where z_α is the value for which a standardized Gaussian is exceeded with probability $\alpha = 5\%$ and $\lceil x \rceil$ denotes a ceiling function that maps x to the smallest integer greater than or 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: (i) a forecasted exceedance that is not observed; (ii) a forecasted exceedance that is observed; and (iii) an observed exceedance that is not forecasted. 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.

$$\mathrm{aFAR} = \left(\frac{\mathrm{Aa}}{\mathrm{Aa}+\mathrm{Ab}}\right)\times 100\%, \tag{C1}$$

$$\mathrm{aH} = \left(\frac{\mathrm{Ab}}{\mathrm{Ab}+\mathrm{Ac}}\right)\times 100\%, \tag{C2}$$

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; and Ac is the number of observed exceedances that are not forecasted within the designated area centered at the monitor location. The aFAR [(C1)] refers to the percentage of false alarms if a forecasted exceedance is not observed within the designated area. The aH [(C2)] refers to the percentage of hits if a forecasted exceedance is observed within the designated area. The aFAR and aH both range from 0% to 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 is close to the threshold T.

$$\mathrm{WSI} = \frac{b+\sum_{1}^{n}\mathrm{IP}}{a+b+c}\times 100\%, \tag{C3}$$

$$\mathrm{IP} = \begin{cases} \dfrac{M-fO}{M-fT}, & \text{if } O<T<M<fO \\ \dfrac{O-fM}{O-fT}, & \text{if } M<T<O<fM \end{cases}, \tag{C4}$$

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 (C4) 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, R., 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 Int. Geoscience and Remote Sensing Symp.*, Fort Worth, TX, Institute of Electrical and Electronics Engineers, 2806–2808, https://doi.org/10.1109/IGARSS.2017.8127581.

Briggs, G., 1969: Plume rise: A critical review. National Technical Information Service Tech. Rep., 81 pp.

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

Cascio, W., 2018: Wildland fire smoke and human health. *Sci. Total Environ.*, **624**, 586–595, https://doi.org/10.1016/j.scitotenv.2017.12.086.

Chin, M., and Coauthors, 2002: Tropospheric aerosol optical thickness from the GOCART model and comparisons with satellite and sun photometer measurements. *J. Atmos. Sci.*, **59**, 461–483, https://doi.org/10.1175/1520-0469(2002)059<0461:TAOTFT>2.0.CO;2.

Darmenov, A., and A. da Silva, 2015: The Quick Fire Emissions Dataset (QFED): Documentation of versions 2.1, 2.2 and 2.4. NASA Global Modeling and Assimilation Office Tech. Rep. Series on Global Modeling and Data Assimilation NASA/TM–2015-104606/Vol. 38, 212 pp., https://ntrs.nasa.gov/api/citations/20180005253/downloads/20180005253.pdf.

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

DelSole, T., 2007: A Bayesian framework for multimodel regression. *J. Climate*, **20**, 2810–2826, https://doi.org/10.1175/JCLI4179.1.

——, and M. Tippett, 2016: Forecast comparison based on random walks. *Mon. Wea. Rev.*, **144**, 615–626, https://doi.org/10.1175/MWR-D-15-0218.1.

Dowell, D., and Coauthors, 2022: The High-Resolution Rapid Refresh (HRRR): An hourly updating convection-allowing forecast model. Part I: Motivation and system description. *Wea. Forecasting*, **37**, 1371–1395, https://doi.org/10.1175/WAF-D-21-0151.1.

Freitas, S., and Coauthors, 2007: Including the sub-grid scale plume rise of vegetation fires in low resolution atmospheric transport models. *Atmos. Chem. Phys.*, **7**, 3385–3398, 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.

Hanna, S. R., 2006: A review of uncertainty and sensitivity analysis of atmospheric transport and dispersion models. Preprints, *28th NATO/CCMS Int. Technical Meeting on Air Pollution Modeling and Its Application*, Leipzig, Germany, NATO/CCMS, 225–237.

Hoerl, A., and R. Kennard, 1970: Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67, https://doi.org/10.2307/1267351.

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. *Atmos. Meas. Tech.*, **4**, 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.*, **120**, 695–701, https://doi.org/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**, 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**, e2020JD032768, https://doi.org/10.1029/2020JD032768.

——, and Coauthors, 2023: Impacts of estimated plume rise on $PM_{2.5}$ 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**, 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**, 969–994, https://doi.org/10.5194/acp-20-969-2020.

Pena, M., and H. van den Dool, 2008: Consolidation of multimodel forecasts by ridge regression: Application to Pacific sea surface temperature. *J. Climate*, **21**, 6521–6538, https://doi.org/10.1175/2008JCLI2226.1.

Randles, C., and Coauthors, 2017: The MERRA-2 aerosol reanalysis, 1980 onward. Part I: System description and data assimilation evaluation. *J. Climate*, **30**, 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**, 144–162, https://doi.org/10.1109/JSTARS.2009.2027443.

Sessions, W. R., 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.

U.S. EPA, 2020: Review of the National Ambient Air Quality Standards for Particulate Matter. *Federal Register*, Vol. 85, No. 244, 82684, https://www.govinfo.gov/content/pkg/FR-2020-12-18/pdf/2020-27125.pdf.

Wooster, M. J., G. Roberts, G. L. W. Perry, and Y. J. Kaufman, 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. *J. Geophys. Res.*, **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). *Quart. J. Roy. Meteor. Soc.*, **145**, 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**, 14 427–14 469, https://doi.org/10.5194/acp-21-14427-2021.

Zhang, J., J. S. Reid, D. L. Westphal, N. L. Baker, and E. J. Hyer, 2008: A system for operational aerosol optical depth data assimilation over global oceans. *J. Geophys. Res.*, **113**, D10208, https://doi.org/10.1029/2007JD009065.

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). Algorithm Theoretical Basis Doc., version 3.1, 30 pp., https://www.ospo.noaa.gov/Products/land/gbbepx/docs/GBBEPx_ATBD.pdf.