JGR Atmospheres

RESEARCH ARTICLE
10.1029/2022JD037298

Key Points:

- Ensemble mean of models provides reliable aerosol optical depth and $PM_{2.5}$ forecasts with stronger correlation, lower bias, and the highest overall *RANKs*
- The ensemble approach provides a well suited exceedance probability forecast during the 2020 Gigafire events
- The multi-model ensemble approach can reduce uncertainties in air quality forecasts and improve model predictability during wildfire events

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:
D. Q. Tong and Y. Li,
qtong@gmu.edu;
yli74@gmu.edu

Citation:
Makkaroon, P., Tong, D. Q., Li, Y., Hyer, E. J., Xian, P., Kondragunta, S., et al. (2023). Development and evaluation of a North America ensemble wildfire air quality forecast: Initial application to the 2020 Western United States "Gigafire". *Journal of Geophysical Research: Atmospheres*, *128*, e2022JD037298. https://doi.org/10.1029/2022JD037298

Received 14 JUN 2022
Accepted 17 OCT 2023

Author Contributions:
Writing – original draft: P. Makkaroon

Development and Evaluation of a North America Ensemble Wildfire Air Quality Forecast: Initial Application to the 2020 Western United States "Gigafire"

P. Makkaroon[1], D. Q. Tong[1,2], Y. Li[1], E. J. Hyer[3], P. Xian[3], S. Kondragunta[4], P. C. Campbell[2,5], Y. Tang[2,5], B. D. Baker[5], M. D. Cohen[5], A. Darmenov[6], A. Lyapustin[6], R. D. Saylor[5], Y. Wang[7], and I. Stajner[8]

[1]Department of Atmospheric, Oceanic and Earth Sciences, George Mason University, Fairfax, VA, USA, [2]Center for Spatial Information Science and Systems, George Mason University, Fairfax, VA, USA, [3]Marine Meteorology Division, Naval Research Laboratory, Monterey, CA, USA, [4]Center for Satellite Applications and Research, NOAA/NESDIS, College Park, MD, USA, [5]NOAA Air Resources Laboratory, College Park, MD, USA, [6]NASA Goddard Space Flight Center, Greenbelt, MD, USA, [7]University of Maryland Baltimore County, Baltimore, MD, USA, [8]National Centers for Environmental Prediction, NOAA National Weather Service, College Park, MD, USA

Abstract Wildfires emit vast amounts of aerosols and trace gases into the atmosphere, exerting myriad effects on air quality, climate, and human health. Ensemble forecasting has been proposed to reduce the large uncertainties in the wildfire air pollution forecast. This study presents the development of a multi-model ensemble (MME) wildfire air pollution forecast over North America. The ensemble members include regional models (GMU-CMAQ, NACC-CMAQ, and HYSPLIT), global models (GEFS-Aerosols, GEOS5, and NAAPS), and global ensemble (ICAP-MME). Performance of the ensemble forecast was evaluated with MAIAC and VIIRS-SNPP retrieved aerosol optical depth (AOD) and AirNow surface $PM_{2.5}$ measurements during the 2020 Western United States "Gigafire" events (August–September 2020). Compared to individual models, the ensemble mean significantly reduced the biases and produced more consistent and reliable forecasts during extreme fire events. For AOD forecasts, the ensemble mean was able to improve model performance, such as increasing the correlation to 0.62 from 0.33 to 0.57 by individual models compared to VIIRS AOD. The ensemble mean also yields the best overall *RANK* (a composite indicator of four statistical metrics) when compared to VIIRS and MAIAC AOD. For the surface $PM_{2.5}$ forecast, the ensemble mean outperformed individual models with the strongest correlation (0.60 vs. 0.43–0.54 by individual models), lowest fractional bias (0.54 vs. 0.55–1.32), highest hit rate (87% vs. 40%–82%), and highest *RANK* (2.83 vs. 2.40–2.81). Finally, the ensemble shows the potential to provide a probability forecast of air quality exceedances. The exceedance probability forecast can be further applied to early warnings of extreme air pollution episodes during large wildfire events.

Plain Language Summary Wildfires are a major source of air pollution emitting large quantities of particles into the *air* that adversely affect human health. Predicting wildfire air pollution, however, is challenging. Ensemble forecasting has been proposed to improve the model predictability. We developed a new multi-model ensemble forecast system of wildfire air pollution over North America, leveraging regional and global atmospheric models by federal agencies and academia. How well the ensemble forecast can predict wildfire pollution was evaluated with observations from satellites and ground monitors. We found that the ensemble mean can significantly reduce the forecast biases and produce more reliable forecasts during extreme wildfire fire events. The ensemble probability forecast of exceedance of the health-based National Ambient Air Quality Standards for fine particles ($PM_{2.5}$) can be further applied to early warnings of severe air pollution episodes during large wildfire events. These findings highlight the potential of the ensemble approach to improve the predictability of air pollution during large wildfires.

1. Introduction

Wildfires are important emission sources that contribute a vast amount of aerosols and trace gases to the atmosphere, leading to hazardous air quality. Exposure to wildfire pollution has been associated with adverse respiratory health issues and premature mortality, which in turn impose substantial economic burdens on society (Fann et al., 2018; Ford et al., 2018; Neumann et al., 2021). Smoke from massive wildfires can generate hazy conditions

© 2023. The Authors.
This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

near the ground, which poses a risk to transportation safety by degrading visibility (Ford et al., 2018; Spracklen et al., 2009). Over the past several decades, the frequency and intensity of both small and large wildfire events in the United States (U.S.) have been rapidly increasing in wildfire-prone areas in the western U.S., such as the Southwest, the Rocky Mountains, the northern Great Plains, and the Pacific Coast (Liu et al., 2013), as a result of climate change from anthropogenic activities causing rising temperatures (Liu et al., 2013; Pierce et al., 2013; Schoennagel et al., 2017). In addition, a sharp increase in the number of small wildfires in the western U.S. is mainly due to human activities, such as changing land cover by expanding cities into wildlands and increasing human ignitions from campfires, powerlines, and vehicles (Hessburg et al., 2019; S. Li & Banerjee, 2021; McClure & Jaffe, 2018; Ryan et al., 2013; Salguero et al., 2020; Stevens-Rumann et al., 2018). The National Interagency Fire Center (NIFC) reported that in 2020 there were 58,950 fires across the U.S., with more than 10 million acres burned (NICC, 2020), and most of the fires took place in the western U.S. In particular, Northern California was significantly impacted and experienced the largest recorded wildfires during the summer of 2020 (California Department of Forestry and Fire Protection [CAL FIRE], 2020).

Many operational forecasting systems have been developed to predict the dispersion and transformation of trace gases and aerosols, with the primary goal of mitigating the health effects of poor air quality (Campbell et al., 2022; Lee et al., 2017; L. Pan et al., 2014; Tang et al., 2015). However, the accuracy of deterministic forecasts from a single model is subject to uncertainties in emission and meteorological input data, as well as the physical and chemical processes of the dispersion or chemical transport models (Delle Monache & Stull, 2003; Kumar et al., 2020; Y. Li et al., 2020). One of the effective ways to improve forecasting performance is the ensemble approach, which can provide probabilistic forecasts by calculating the mean from either multiple models or varying inputs in a single model (Delle Monache, Deng, et al., 2006; Delle Monache, Nipen, et al., 2006; Delle Monache et al., 2008, 2020; Delle Monache & Stull, 2003; Y. Li et al., 2020; Petersen et al., 2019; Solazzo et al., 2012; Xian et al., 2019). The major advantage of the ensemble forecast over a single model forecast is that it can reduce the biases in the forecasts of ensemble members by averaging them out, whereas the uncertainties in ensemble forecasts can also be determined from the spreads of ensemble members.

This work presents the development of a multi-model ensemble (MME) wildfire forecasting system and its application to air quality prediction during the 2020 Gigafire in the Western U.S. The motivation for developing aerosol MME consensus is based on numerical weather prediction studies that have shown the usefulness of ensemble-mean-based predictions in understanding systematic errors arising from models' imperfect nature and the sensitivity of models to initial conditions. For example, multi-model consensuses are found on average to produce more accurate forecasts of cyclone track and intensity than the individual model members (e.g., Goerss et al., 2004; Sampson et al., 2008). In previous studies, a regional single-model ensemble was created to predict surface $PM_{2.5}$ during the 2018 California Camp Fire event using the NOAA HYSPLIT dispersion model at 0.1-degree resolution (Y. Li et al., 2020), and the International Cooperative for Aerosol Prediction (ICAP) MME has been established to provide AOD forecasts globally (Xian et al., 2019).

The new MME forecast system is based on seven real-time operational forecast systems including three regional models, three global models, and one global ensemble. The regional systems include: (a) the George Mason University-Community Multiscale Air Quality (GMU-CMAQ) model (Y. Li et al., 2021), (b) the National Oceanic and Atmospheric Administration-U.S. Environmental Protection Agency (NOAA-EPA) Atmosphere-Chemistry Coupler-Community Multiscale Air Quality model (NACC-CMAQ) (Campbell et al., 2022), and (c) the NOAA Hybrid Single-Particle Lagrangian Integrated Trajectory (HYSPLIT) model (Rolph et al., 2009). The three global models are: (a) Global Ensemble Forecast System Aerosols (GEFS-Aerosols) (Hamill, Whitaker, Fiorino, & Benjamin, 2011), (b) NASA Goddard Earth Observing System (GEOS, version 5) (Buchard et al., 2017; Randles et al., 2017), and (c) Navy Aerosol Analysis and Prediction System (NAAPS) (Lynch et al., 2016). Finally, the International Cooperative for Aerosol Prediction Multi-Model aerosol forecasting Ensemble (ICAP-MME) is a global ensemble mean produced from nine comprehensive global speciated aerosol and/or dust models (Xian et al., 2019). The ICAP-MME includes predictions from some global models not listed above but not from any regional models. The goal of this study is to improve real-time wildfire air quality forecasts with ensemble forecasts with available operational and research real-time forecasts that are already implemented at various agencies.

The goal of this study is to improve real-time wildfire air quality forecasts with ensemble forecasts with available operational and research real-time forecasts that are already implemented at various agencies. This study advances wildfire air quality forecasting over the North America region in several ways. A new MME forecast system has been developed using both global and regional forecasting models operated by U.S. federal agencies and university research centers. The new system was applied to predict AOD and surface fine particulate matter ($PM_{2.5}$) during the unprecedented 2020 Western U.S. Gigafire to test its forecasting performance for extremely large wildfire events. The capability to predict Gigafire will better prepare society for future events in a warming climate. Furthermore, we developed a probability forecast of $PM_{2.5}$ exceedance based on the results of the ensemble members. As air quality forecasts are often used to issue early warnings to the public, it is crucial for the ensemble to produce a reliable forecast of $PM_{2.5}$ exceedances (binary prediction) of health-based National Ambient Air Quality Standards (NAAQS) (U.S. EPA, 2020b) during wildfire events. The MME members are introduced in Section 2, and the evaluation of the ensemble forecasts and results for the $PM_{2.5}$ exceedance probability forecast are discussed in Section 3. We conclude in Section 4.

2. Materials and Methods

2.1. The 2020 Gigafire in the Western United States

In 2020, the western U.S. experienced multiple complex wildfires, leading to a “Gigafire” that burned over 10.2 million acres (NIFC, 2020) across California, Oregon, and Washington. A significant proportion of burned areas were a consequence of California's largest complex fire ever, known as the “August Complex Fire” during August–September 2020, resulting from lightning strikes (CAL FIRE, 2020) driven by a heatwave, severe drought (Guirguis et al., 2018; Hulley et al., 2020; Pathak et al., 2018), and daytime southerly winds (Varga et al., 2022). Smoke from the wildfires spread across the western U.S. leading to hazardous air quality predominantly in California, Oregon, Washington, Idaho, and Nevada. Figure 1a displays high observed $PM_{2.5}$ concentrations (>250 μg/m^3; maroon red) from AirNow sites, mainly in the western U.S., on 12 September 2020, when the fires were very intense. Also, high daily $PM_{2.5}$ concentrations above the daily NAAQS for $PM_{2.5}$ (>35 μg/m^3) were recorded at many AirNow monitoring sites across the U.S. between the middle of August–September, primarily over California, Oregon, and Washington (Y. Li et al., 2021) as shown in Figure 1b. Consequently, our study will focus on AOD and $PM_{2.5}$ simulations during the Gigafire events from August to September 2020.

2.2. Description of Ensemble Members

According to previous studies, the challenge of the wildfire air quality forecast stems in part from the uncertainties in fire emission and the estimation of the plume height (Y. Li et al., 2020, 2023). The forecast models here used to create the ensemble cover a various range of emission data sets, as well as different plume rise schemes. In this section, each of the seven participating numerical atmospheric models included in the ensemble will be described. Model configurations are shown in Table S1 in Supporting Information S1.

2.2.1. GMU-CMAQ

GMU-CMAQ is a research forecasting system run by the air quality group of George Mason University (GMU) to provide daily air quality forecasts across the U.S. for the general public (Y. Li et al., 2021). The model uses meteorological fields derived from the Weather Research and Forecasting model version 4.2 (WRFv4.2) (Skamarock et al., 2019) with the meteorological initial and boundary from the National Centers for Environmental Prediction (NCEP) operational Global Forecast System (GFS) to drive the offline CMAQ model version 5.3.1 (CMAQv5.3.1) (U.S. EPA, 2020a), and uses biomass burning (BB) emission data from the Global BB Emissions Product (GBBEPx; X. Zhang et al., 2012, 2014, 2019) calculated using averaged fire emissions from Terra Moderate Resolution Imaging Spectroradiometer (MODIS) fire radiative power (FRP), Aqua MODIS FRP, VIIRS-SNPP FRP and Joint Polar Satellite System (JPSS) 1 VIIRS FRP. The anthropogenic emission data is taken from the U.S. EPA 2016 National Emissions Inventory Collaborative version 1 (2016v1) Emission Modeling Platform, which is generated by the Sparse Matrix Operator Kennel Emissions model version 4.7 (Houyoux et al., 2000) using the base year of the emission inventory taken from the 2016v1 Emission Modeling Platform (Eyth et al., 2020). The wildfire smoke plumes are calculated using the Sofiev et al. (2012) plume rise algorithm. GMU-CMAQ provides hourly experimental AOD and $PM_{2.5}$ concentration forecasts on a horizontal resolution of 12 × 12 km over the CONUS with each day's forecast initialized at 18:00 UTC on the previous day.

2.2.2. NACC-CMAQ

The NACC-CMAQ is a model currently being used in NOAA's operational National Air Quality Forecasting Capability. It used a meteorological preprocessor adapted from the EPA's Meteorology Interface Processor version

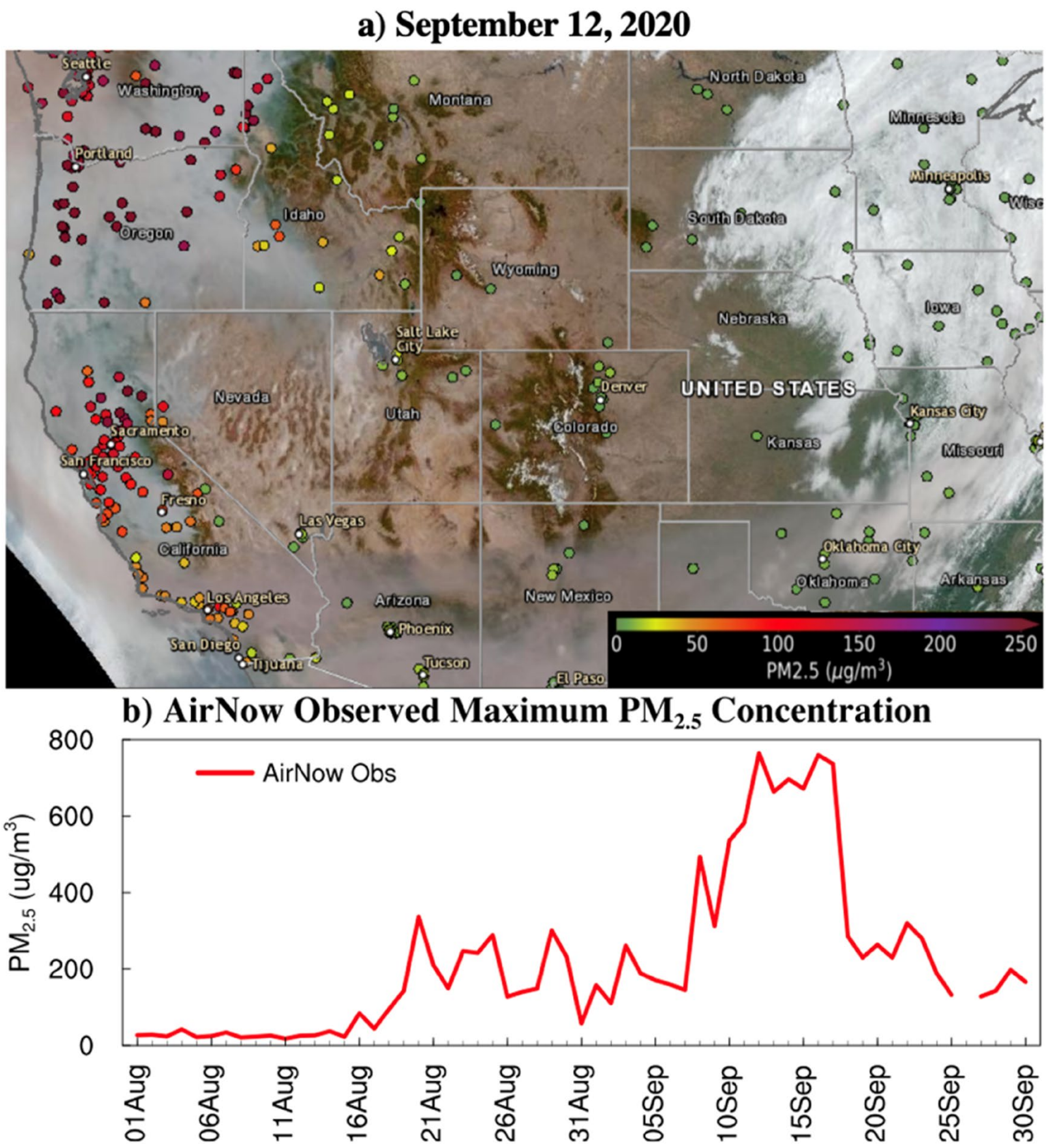

Figure 1. (a) VIIRS-SNPP true color imagery overlaid by daily mean $PM_{2.5}$ observations measured by AirNow sites (circles) on 12 September 2020, from NOAA AerosolWatch. (b) The time series plot of daily maximum $PM_{2.5}$ concentrations measured by all AirNow sites across the Contiguous United States during the Gigafire events from August to September 2020.

5 (e.g., NACC version 1.3.2; https://zenodo.org/record/5507489#.YmvzsejMKUk, last access 29 April 2022) that ingests the outputs from NOAA's latest operational Finite Volume Cubed-Sphere Global Forecast System version 16 to prepare the meteorology files that are used within the CMAQ Modeling System (Campbell et al., 2022). Emission input data sets are very similar to GMU-CMAQ and include GBBEPx for BB emissions, NEI 2016v1 for anthropogenic emissions, and Biogenic Emission Inventory System version 3.6.1 (BEISv3.6.1; Vukovich & Pierce, 2002; Schwede et al., 2005) with the Biogenic Emission Landuse Data set version 5 for biogenic volatile organic carbon (BVOC) emissions. The wildfire smoke plumes are computed using the Briggs (1969) plume rise algorithm. NACC-CMAQ uses meteorology and emission inputs together with lateral boundary conditions from NOAA's operational GEFS-Aerosols model to account for long-range transport of air pollution for dust and smoke to provide hourly AOD and $PM_{2.5}$ forecasts at a horizontal resolution of 12 × 12 km (same as GMU-CMAQ) with each day's forecast initialized at 12:00 UTC of the previous day over CONUS.

2.2.3. HYSPLIT

HYSPLIT is a widely used atmospheric transport and dispersion model developed at the NOAA Air Resources Laboratory (ARL) (Stein et al., 2015). The model uses a plume-following coordinate system and is typically used to determine the atmospheric transport, dispersion, deposition, and chemical transformation of pollutants over regional and global domains. Since 2007, it has been employed in NOAA's Smoke Forecasting System using fire locations from satellite data and BB data based on vegetation cover from the bottom-up, fuel-based Blue Sky modeling system developed by the U.S. Forest Service (Rolph et al., 2009; Stein et al., 2009). The HYSPLIT-based Smoke Forecasting System configuration used in this study combines meteorology inputs from the NCEP North American Mesoscale 12-km model, fire locations from the NOAA NESDIS Hazard Mapping System (HMS), and emissions from fire locations from the U.S. Forest Service (USFS) BlueSky framework (Larkin et al., 2009). The wildfire smoke plumes are computed using the Briggs (1969) plume rise scheme to simulate hourly AOD and $PM_{2.5}$ concentration forecasts at a horizontal resolution of 0.15° × 0.15° with each day's forecast initialized at 00:00 UTC on the previous day over CONUS.

2.2.4. GEFS-Aerosols

NOAA's GEFS-Aerosols is a global atmospheric composition model established at the NCEP in collaboration with the NOAA Global Systems Laboratory, NOAA Chemical Sciences Laboratory, and NOAA/ARL. The GEFS-Aerosols version 1 model used here provides aerosol and atmospheric composition forecasts using FV3-based GFSv15 meteorology coupled to NASA GOCART aerosol model component using the National Unified Operational Prediction Capability Layer (Theurich et al., 2016), which is the current and future foundation of NOAA's Unified Forecast System modeling framework (Hamill, Whitaker, Fiorino, & Benjamin, 2011; Hamill, Whitaker, Kleist, et al., 2011; L. Zhang et al., 2022; Zhou et al., 2022). The operational GEFS-Aerosols model currently uses BB emission data from GBBEPx and global anthropogenic emission data from the Community Emission Data System in 2014 for gaseous emissions and the Hemisphere Transport of Air Pollution (HTAP) version 2 for primary aerosol emissions. Wildfire smoke plumes are calculated using a one-dimension (1-D) time-dependent cloud module from High-Resolution Rapid Refresh (HRRR)-Smoke model (Freitas et al., 2007). This study employed GEFS-Aerosols global AOD and $PM_{2.5}$ forecasts at a horizontal resolution of 0.25° × 0.25° and initialized each day at 00:00 UTC.

2.2.5. GEOS

The Goddard Earth Observing System (GEOS) is a modular modeling system that can be configured to conduct basic research and to support a range of applications related to Earth Science, including short-range weather prediction, field mission support, subseasonal-to-seasonal forecasting, and generation of multidecadal reanalysis. The GEOS system is developed by NASA's Global Modeling and Assimilation Office. This study used the GEOS Forward Processing system (GEOS-FP, version 5.27.1) which generates analyses, assimilation products, and 10-day forecasts in near-real time. GEOS-FP is built around the GEOS AGCM, the GEOS atmospheric data assimilation system (hybrid–4DEnVar ADAS), and aerosol assimilation (Randles et al., 2017). Aerosols are an integral component of the model physics (Buchard et al., 2017) and are simulated with the Goddard Chemistry, Aerosol, Radiation, and Transport model (GOCART; Chin et al., 2002; Colarco et al., 2010). Fire emissions come from the Quick Fire Emissions Data set (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). Anthropogenic emissions are from the Emissions Database for Global Atmospheric Research and Hemispheric Transport of Air Pollution (HTAP) inventories. BVOC emissions are from the Model of Emissions of Gases and Aerosols from Nature (MEGAN). This study used GEOS global forecast of hourly AOD values and $PM_{2.5}$ concentrations on a horizontal resolution of 0.25° × 0.3125° and initialized each day at 00:00 UTC.

2.2.6. ICAP-MME

Established in 2010, ICAP aims to promote community development of global aerosol observations, data assimilation, and prediction technologies to support operational aerosol forecasting (Benedetti et al., 2011; Colarco et al., 2014; Reid et al., 2011). The ICAP-MME (Sessions et al., 2015; Xian et al., 2019) is a global multi-model aerosol forecasting ensemble consensus (currently, only AOD product is available) maintained by the Marine Meteorology Division of the Naval Research Laboratory (NRL), which provides a testbed of probabilistic aerosol forecasts. ICAP-MME is generated by combining nine global aerosol models: the European Center for Medium-range Weather Forecasts-Monitoring Atmospheric Composition and Climate model (ECMWF) under Copernicus Atmosphere Monitoring Service (CAMS, former MACC), GEOS, NAAPS, Japan Meteorological

Agency Model of Aerosol Species in the Global Atmosphere, NOAA Environmental Modeling System GFS Aerosol Component (NGAC), Mětěo-France Modělě de Chimie Atmospherique ã Grande Echelle, and Finnish Meteorological Institute System for Integrated modeLing of Atmospheric coMposition, the Barcelona Supercomputing Center Chemical Transport Model, embedded in the Multiscale Online Nonhydrostatic AtmospheRe CHemistry and the UK Met Office models. These models have different underlying meteorological fields, emissions, microphysics, and chemistry, and several include assimilation of satellite aerosol data, though using diverse processing methods and assimilation techniques. The horizontal and vertical resolutions of these models range from 0.25° × 0.31° and 72 vertical layers to 1.4° × 1° and 24 layers. As a result, ICAP-MME is driven by the independent operation/quasi-operational meteorological data set, emission inputs, and plume rise algorithms generated by each of the member organizations. This study utilized ICAP-MME global 6-hr AOD at 550 nm on a horizontal resolution of 1° × 1° and initialized each day at 00:00 UTC.

2.2.7. NAAPS

NAAPS is developed at the Marine Meteorology Division of the NRL and provides an operational forecast of 3D atmospheric anthropogenic fine and biogenic fine aerosols, BB smoke, dust, and sea salt concentrations (Lynch et al., 2016). The current NAAPS is driven by global meteorological fields from the Navy Global Environmental Model, an operational global weather prediction system developed by the U.S. Navy (Hogan et al., 2014). NAAPS uses a BB smoke source from the Fire Locating and Modeling of Burning Emissions inventory, which is based on near-real time MODIS fire hotspot data (Reid et al., 2009). The wildfire smoke plumes are distributed uniformly through the bottom 4 layers. The NAAPS analysis is constrained by the assimilation of MODIS AOD (Hyer et al., 2011; J. Zhang et al., 2008). This study employed the NAAPS global 3-hourly AOD and surface $PM_{2.5}$ concentrations at a horizontal resolution of 0.333° × 0.333° and initialized each day at 00:00 UTC.

2.3. Description of Observations

The performance of the ensemble in forecasting wildfire air pollution is verified with satellite AOD observations from the MODIS aboard Terra and Aqua and Visible Infrared Imaging Radiometer Suite onboard the Suomi National Polar-orbiting Partnership (SNPP) (VIIRS-SNPP) satellite and with ground $PM_{2.5}$ ground observations from the EPA AirNow network.

2.3.1. AirNow $PM_{2.5}$

Hourly $PM_{2.5}$ observations were obtained from the U.S. EPA AirNow network (https://www.AirNow.gov). The AirNow data sets are acquired from a variety of monitoring data collected by AirNow and its partners, such as the EPA, NOAA, National Park Service, NASA, Centers for Disease Control, and tribal, state, and local air quality agencies, using a federal reference or equivalent monitoring methods approved by EPA. In this study, hourly $PM_{2.5}$ concentrations, starting from 12:00 UTC of the current day to 11:00 UTC the next day, were derived from each AirNow site and averaged into a daily value for each site's location. This should be noted that the same time period was applied to calculate daily $PM_{2.5}$ concentrations simulated by the individual models and ensemble.

2.3.2. MAIAC AOD

MAIAC algorithm is designed to work with the time series and spatial analyses of the MODIS L1B data, which are gridded to a fixed 1 km grid resolution to observe the same grid cell over time, resulting in an improvement in the accuracy of aerosol retrievals, atmospheric correction, and cloud detection (Lyapustin et al., 2012, 2018; Lyapustin, Martonchik, et al., 2011; Lyapustin, Wang, et al., 2011). MAIAC provides plume injection height in the same suite of MCD19A2 products (Lyapustin et al., 2020). In addition to standard MODIS calibration, in Collection 6 and beyond, MAIAC applies a residual de-trending of both MODIS Terra and Aqua sensors, along with polarization correction of MODIS Terra and cross-calibration of Terra to Aqua (Lyapustin et al., 2014). This allows MAIAC to process MODIS Terra and Aqua jointly as a single sensor. This study used mean daily global 1 km MAIAC AOD at 550 nm averaged from all orbits available for the CONUS. MAIAC data were provided by NASA Goddard Space Flight Center. Note that while several of the input models are constrained by the assimilation of MODIS data, none in this study use the MAIAC data, so it is at least partially independent from all models.

2.3.3. VIIRS-SNPP AOD

VIIRS-SNPP AOD product was acquired from the VIIRS instrument carried onboard the Suomi National Polar-orbiting Partnership (SNPP) (Cao et al., 2013, 2014; Uprety et al., 2013), which is a part of the JPSS. The

VIIRS instrument was initially developed based on the previous series of measurements on NOAA satellites and MODIS on Terra and Aqua satellites through the cooperation of NASA and NOAA (Levy et al., 2013, 2015). The instrument provides improved operational environmental monitoring and sensor data records for aerosol products through a short-wave infrared spanning from 0.412 to 2.25 microns to support NASA's EOS and NOAA's polar-orbiting operational environmental satellite system. VIIRS-SNPP observes the entire Earth's surface twice each day. It passes the equator at approximately 13:30 local time (LST). In this study, we used VIIRS-SNPP Level 3 enhanced Dark Target over dark and bright surfaces AOD products at 550 nm (H. Zhang et al., 2016) with a fixed grid resolution at 0.1° as provided by NOAA. None of the models in this study used VIIRS AOD data in data assimilation during this study period.

2.4. Multi-Model Ensemble Forecast

We created the multi-model ensembles for AOD and $PM_{2.5}$ during the 2020 Gigafire events (August to September 2020). For the AOD ensemble, we used the unweighted arithmetic mean value of AOD simulated by GMU-CMAQ, NACC-CMAQ, HYSPLIT, ICAP-MME, GEFS-Aerosols, GEOS, and NAAPS models. For the $PM_{2.5}$ ensemble, we used the unweighted arithmetic mean value of $PM_{2.5}$ concentrations simulated by GMU-CMAQ, NACC-CMAQ, HYSPLIT, GEFS-Aerosols, GEOS, and NAAPS models. Data from all models were interpolated to a unified horizontal grid of 12 × 12 km before calculating the ensemble mean. The ensembles provide 24-hr forecasts of AOD and $PM_{2.5}$ across the CONUS.

Note the ensemble forecasts calculate the mean of the member models, without applying weighting factors to these models. The use of unweighted arithmetic mean value may not bring out the full potential of the ensemble forecasting system, as a weighted ensemble system may yield better performance than the unweighted ensemble system. We constructed two additional ensembles, one using a weighted ensemble method and the other with the ensemble median for AOD and $PM_{2.5}$. Compared to the unweighted ensemble mean, the weighted ensemble shows mixed results in the model performance. While reducing biases, including mean bias (MB) and normalized mean bias (NMB), the weighted ensemble did not improve the model performance in terms of root mean square error (RMSE), correlation, or errors. In addition, the Gigafire event studied here is unprecedented in many ways, including the burned areas, emission amounts, fire intensity, and burning duration. Because of the exceptionalness of this event, there is no historic training data available to derive weighting factors, which makes it difficult to derive the weighting factors needed for the weighted ensemble method.

2.5. Ensemble Probability of $PM_{2.5}$ Exceedance Forecast

The GMU-CMAQ, NACC-CMAQ, HYSPLIT, GEFS-Aerosols, GEOS, and NAAPS were used to create the ensemble probability of the $PM_{2.5}$ exceedance forecast. The probability was calculated using Equation 1 based on the number of models that forecast $PM_{2.5}$ exceedances (daily average concentrations larger than 35 $\mu g/m^3$) during the Gigafire events. The probability result ranges from 0% (none of the models forecasting the exceedances; exceedances are very unlikely to occur) to 100% (all models forecasting the exceedances; exceedances are very likely to occur):

$$P(A) = \frac{\text{Number of models that forecast the exceedances}}{\text{Total number of models}} \times 100\% \tag{1}$$

where $P(A)$ is the probability of event A.

2.6. Model Evaluation

The AOD and surface $PM_{2.5}$ concentrations simulated by the ensemble mean and individual models are evaluated with satellite and ground measurements, respectively. The performance of AOD prediction by individual models and ensemble mean was evaluated against VIIRS and MAIAC AOD. For $PM_{2.5}$, the performance of $PM_{2.5}$ forecasts was verified by comparing model simulations against daily average $PM_{2.5}$ observations from the EPA AirNow ground network. In both cases, the simulations by individual models and the ensemble mean were compared with observations at the nearest satellite retrieval or monitoring site. Any grids containing missing model data and/or unqualified observations data were excluded from the calculation.

A suite of statistical metrics, including RMSE, correlation (CORR), absolute fractional bias (FB), MB, and mean error (ME) were calculated using Equations 2–6:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=0}^{N}(M_i - O_i)^2} \tag{2}$$

$$\mathrm{CORR} = \frac{N\sum_{i=0}^{N} O_i M_i - \sum_{i=0}^{N} O_i \sum_{i=0}^{N} M_i}{\sqrt{N\sum_{i=0}^{N} O_i^2 - \left(\sum_{i=0}^{N} O_i\right)^2}\sqrt{N\sum_{i=0}^{N} M_i^2 - \left(\sum_{i=0}^{N} M_i\right)^2}} \tag{3}$$

$$\mathrm{FB} = 2 \times \frac{\sum_{i=0}^{N} |O_i - M_i|}{\sum_{i=0}^{N} |O_i + M_i|} \tag{4}$$

$$\mathrm{MB} = \frac{1}{N}\sum_{i=0}^{N}(M_i - O_i) \tag{5}$$

$$\mathrm{ME} = \frac{1}{N}\sum_{i=0}^{N}|M_i - O_i| \tag{6}$$

where M_i represents the ith model forecast, O_i is the ith observation, and N is the total number of observations and time-space matched prediction during the study periods.

The statistical metrics listed above are limited in their ability to evaluate model performance to meet certain thresholds, such as the exceedance of the NAAQS for $PM_{2.5}$. In addition, exceedance forecasts are typically generated for an area, such as a metropolitan or zip code. When an exceedance is predicted to occur at any grid within this area, an early warning will be issued for the whole area. Therefore, we employed two categorical metrics: the area hit rate (*aH*) and the area false alarm ratio (*a*FAR) following Kang et al. (2007). These two metrics will supplementarily measure the forecasting performance of individual models, the ensemble mean, and the ensemble probability in predicting daily $PM_{2.5}$ exceedances (24-hr average $PM_{2.5}$ concentrations greater than 35 μg/m^3 based on NAAQS) to reflect the spatial uncertainties of the model forecast. These two metrics were calculated based on paired observed and predicted $PM_{2.5}$ exceedances by considering four possible scenarios: (a) a forecasted exceedance that is not observed (false alarm); (b) a forecasted exceedance that is observed (hit); (c) no exceedance is forecasted or observed; (d) an observed exceedance that is not forecasted. The *aH* and *a*FAR values are determined by matching observed and forecasted exceedances within a designated area surrounding the observation locations. In the present study, we used an area of 0.5° × 0.5° centered at each AirNow site's location. The area hit rate *aH* (Equation 7) refers to the percentage of hits if a forecasted exceedance is observed within the designated area (*A*). The *a*FAR (Equation 8) refers to the percentage of false alarms if a forecasted exceedance is not observed within the designated area. The *aH* and *a*FAR both range from 0% to 100%:

$$aH = \left(\frac{Ab}{Ab + Ad}\right) \times 100\% \tag{7}$$

$$a\mathrm{FAR} = \left(\frac{Aa}{Aa + Ab}\right) \times 100\% \tag{8}$$

where *Aa* is the number of forecasted exceedances that are not observed within a designated area (false positives), *Ab* is the number of forecasted exceedances that are observed within a designated area (hits), and *Ad* is the number of observed exceedances that were not forecasted within a designated area (misses). The *aH* is based on total observed exceedances while the *a*FAR is based on total forecasted exceedances. If a model performs well, the misses (*Ad*) 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 *a*FAR.

Table 1

Overall Ensemble Mean and Individual Model Performances in Forecasting AOD Values and Surface $PM_{2.5}$ Concentrations During the 2020 Gigafire Events (August–September 2020) Based on Seven Statistical Metrics: RMSE, CORR, MB, ME, FB, aH and aFAR, and Overall Rating (RANK)

Cases	Models	RMSE	CORR	MB	ME	FB	%aH	%aFAR	RANK
AOD simulations compared against VIIRS retrievals	Model-1	0.31	0.57	−0.10	0.17	0.66			1.45
	Model-2	0.34	0.49	**−0.17**	0.20	0.85			1.32
	Model-3	0.39	**0.33**	−0.16	0.24	**1.37**			**0.98**
	Model-4	0.32	0.48	−0.12	0.18	0.72			1.38
	Model-5	**0.51**	0.46	0.10	**0.28**	0.79			1.33
	Model-6	0.34	0.52	−0.17	0.20	0.93			1.29
	Model-7	0.31	0.53	**−0.08**	0.17	**0.61**			1.46
	Ensemble Mean	**0.28**	**0.62**	−0.10	**0.16**	0.65			**1.48**
AOD simulations compared against MAIAC retrievals	Model-1	0.25	0.62	−0.07	0.14	0.62			1.50
	Model-2	0.29	0.53	**−0.14**	0.16	0.77			1.38
	Model-3	0.34	**0.40**	−0.12	0.21	**1.24**			**1.08**
	Model-4	0.27	0.53	−0.08	0.15	0.65			1.44
	Model-5	**0.49**	0.49	0.14	**0.28**	0.80			1.35
	Model-6	0.28	0.55	−0.12	0.16	0.82			1.36
	Model-7	0.25	0.59	**−0.05**	0.13	**0.53**			1.53
	Ensemble Mean	**0.22**	**0.67**	−0.06	**0.12**	0.56			**1.55**
$PM_{2.5}$ simulations compared against AirNow observations	Model-1	25.61	0.54	**3.37**	9.97	0.55	69.08	44.29	2.81
	Model-2	**16.96**	0.49	−4.89	**8.02**	0.59	**39.67**	**23.73**	2.74
	Model-3	20.57	**0.43**	−4.65	10.65	**1.32**	71.71	45.42	2.40
	Model-4	**51.96**	0.49	**19.80**	**23.00**	0.89	79.16	**75.57**	**2.38**
	Model-5	51.94	0.44	13.02	18.42	0.77	81.00	68.68	2.50
	Model-7	39.98	0.51	12.30	15.91	0.70	80.79	62.51	2.63
	Ensemble Mean	25.93	**0.60**	7.40	11.16	**0.54**	**86.85**	60.52	**2.83**

Note. The best results of each statistical metric and RANK are highlighted in red and the poorest results are highlighted in blue.

The overall rating (RANK) was used to determine the comprehensive forecasting performances of individual models and ensemble mean during the study period (Draxler, 2006; Y. Li et al., 2020). In the case of $PM_{2.5}$ evaluation, the RANK was derived from the sum of the normalized CORR, FB, aH, and aFAR (Equation 9). In the case of AOD evaluation, the RANK was calculated using the sum of the normalized CORR and FB (Equation 10). $PM_{2.5}$ RANK ranges from 0 to 4 (from worst to best), while AOD RANK ranges from 0 to 2:

$$\mathrm{RANK}_{\mathrm{PM}_{2.5}} = \frac{\mathrm{CORR}+1}{2} + \left(1 - \frac{\mathrm{FB}}{2}\right) + \frac{aH}{100\%} + \left(1 - \frac{a\mathrm{FAR}}{100\%}\right) \tag{9}$$

$$\mathrm{RANK}_{\mathrm{AOD}} = \frac{\mathrm{CORR}+1}{2} + \left(1 - \frac{\mathrm{FB}}{2}\right) \tag{10}$$

3. Results and Discussions

In this section, the MME mean was evaluated with ground and satellite observations during the 2020 Western U.S. Gigafire events (August–September 2020). The forecasting performance of the ensemble mean was also compared with individual models to assess whether the ensemble mean can outperform the top performers among these members. The evaluation results are based on the average of daily statistical metrics and overall rating (RANK) that were calculated from every grid with complete model simulation and observation data sets over the study period (August–September 2020), as shown in Table 1. High values of CORR, aH, and RANK, and low values of RMSE, MB, ME, aFAR, and FB indicate good agreement between model forecasts and observations.

In Sections 3.1, 3.2, and 3.4, we decided to show the forecasts and observations on 22 August 2020, as the case study due to better data completeness of both model simulation and observation data compared to other days within the study period.

3.1. Performance of Ensemble AOD Forecasting

First, we will evaluate the AOD simulations against VIIRS and MAIAC AOD retrievals. Overall, the ensemble forecasts of AOD are promising, where major plume characteristics are well captured across the northwestern U.S. but not so over the central U.S. For instance, contour maps of AOD forecasts, VIIRS AOD, and MAIAC AOD retrievals on 22 August 2020, in Figures 2a–2g and 3a–3g, indicate that the AOD simulations from all the models, Model-1 to 7, underestimated AOD values over the western U.S., while Model-5 overestimated AOD primarily in California (Figures 2c, 2e, 3c, and 3e). In comparison, the ensemble mean slightly overestimated AOD over Northern California and underestimated it over Montana, Wyoming, Colorado, Nebraska, and Kansas, where complex geographic formations, such as the Colorado Plateau-Central Rockies areas, are located (Figures 2h and 3h). However, the majority of areas showing high AOD in the ensemble forecast match fairly well with the satellite observations.

Figures 5 and 6 show the time series of five statistical metrics for AOD evaluation during August-September 2020 against the VIIRS and MODIS MAIAC, respectively. By all of these metrics, the ensemble mean shows better performance compared to individual models. Compared to individual models (dashed lines), the ensemble mean (solid black line) shows a consistently higher correlation (CORR), especially in the mid-August and mid-September when the wildfires were most active (Figures 5 and 6a), and comparatively low values of root mean square error (RMSE; Figures 5 and 6b), mean bias (MB; Figures 5 and 6c), mean error (ME; Figures 5 and 6d), and fractional bias (FB; Figures 5 and 6e). The ensemble mean and most individual models slightly underestimated AOD almost the entire period, especially during the peak fire season (in the middle of September 2020), leading to relatively high negative MB values during this time (Figures 5 and 6b). Similar to previous results, Model-5 overestimated the AOD with relatively high positive MB values during the same period. The errors in model simulations of AOD throughout the extreme fire period were also demonstrated by highly varying values of RMSE, ME, and FB in mid-August and mid-September, as shown in Figures 5, 6b, 6d, and 6e.

Compared with VIIRS AOD (Table 1), the ensemble mean increased correlation and greatly reduced bias and error, as indicated by the highest CORR (0.62), the lowest RMSE (0.28) and ME (0.16), and the second lowest in MB (−0.10 μg/m^3) and FB (0.65), leading to ranking the first place in overall rating (RANK; 1.48). Similar to the comparison with MAIAC AOD, the ensemble mean significantly increased correlation and reduced bias and error suggested by the highest CORR (0.67), the lowest RMSE (0.22), and ME (0.12), and the second lowest in MB (−0.06 μg/m^3) and FB (0.56), resulting in ranking the first place in overall rating (RANK; 1.55). These bias and error values are closer to zero relative to most individual models, meaning that the ensemble mean significantly reduces bias and uncertainties in AOD forecasting.

All the results suggest that the ensemble forecast is capable of reducing the bias in AOD prediction, especially during extreme wildfire events. Furthermore, the ensemble mean successfully produces more statistically consistent and reliable forecasts of AOD during the wildfires relative to the forecasts provided by individual models.

3.2. Ensemble Forecast Performance for Surface $PM_{2.5}$ Concentration

Next, we will evaluate the $PM_{2.5}$ simulations against AirNow ground observations. The results show that the MME mean performs fairly well in forecasting surface $PM_{2.5}$ during extreme wildfires, such as on 22 August 2020 (Figure 4). Figures 4a, 4c, 4d, 4e, and 4f show that Model-1, 3, 4, 5, and 7 overestimated $PM_{2.5}$ concentrations considerably in the western U.S. and to a less extent in the Central and Southern U.S. In contrast, the ensemble mean overestimated $PM_{2.5}$ simulations in Northern California (Figure 4g). However, the extremely high $PM_{2.5}$ concentrations simulated by the ensemble mean are located over the areas that are in fairly good agreement with the AirNow ground observations (Figure 4h).

Figure 7 shows the time series of statistical metrics of surface $PM_{2.5}$ evaluation from August to September 2020. The ensemble mean shows better performance compared to individual models when compared to surface $PM_{2.5}$ observations from the AirNow network. In Figure 7b, the ensemble mean (solid black line) shows consistently

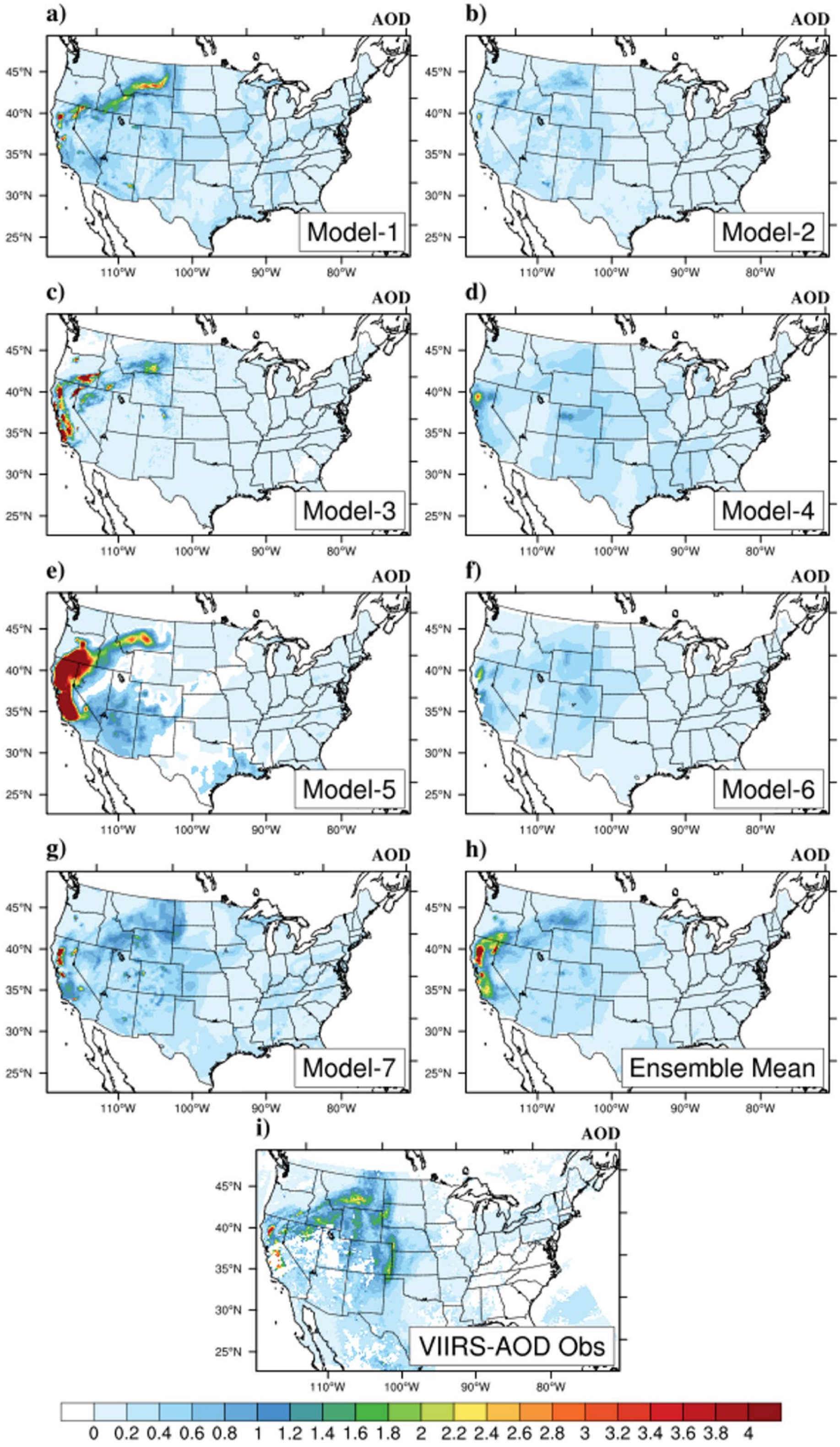

Figure 2. (a–g) Aerosol optical depth (AOD) predicted by seven individual models and (h) the ensemble mean, compared with (i) VIIRS AOD retrievals on 22 August 2020 (during the 2020 Gigafire events). Gaps in the figures are satellite observations that did not pass quality control, under cloud cover, or missing model simulations, which were not used in the calculation.

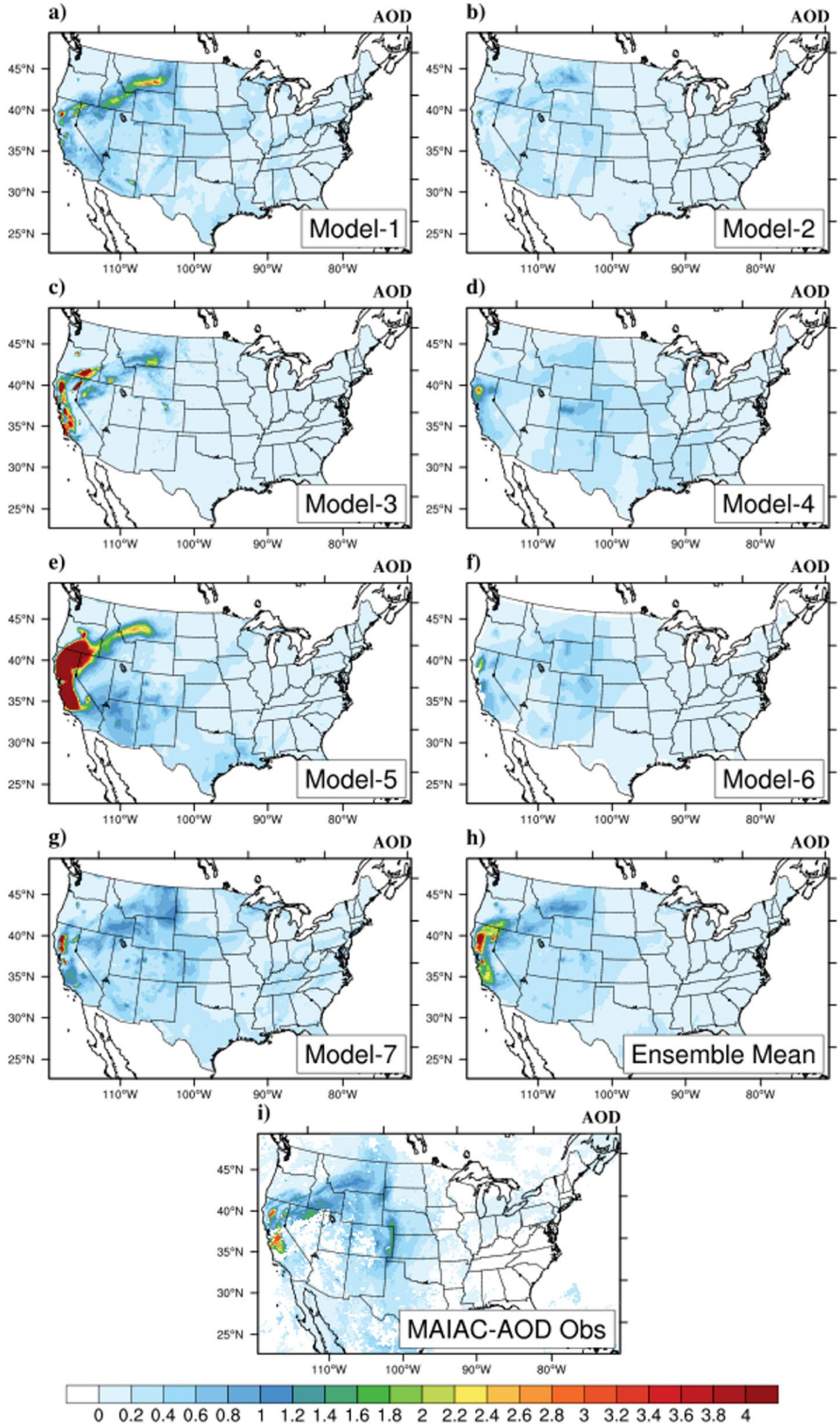

Figure 3. (a–g) Aerosol optical depth (AOD) predicted by seven individual models and (h) the ensemble mean, compared with (i) MAIAC AOD retrievals on 22 August 2020 (during the 2020 Gigafire events). Gaps in the figures are satellite observations that did not pass quality control and missing model simulations, which were not used in the calculation.

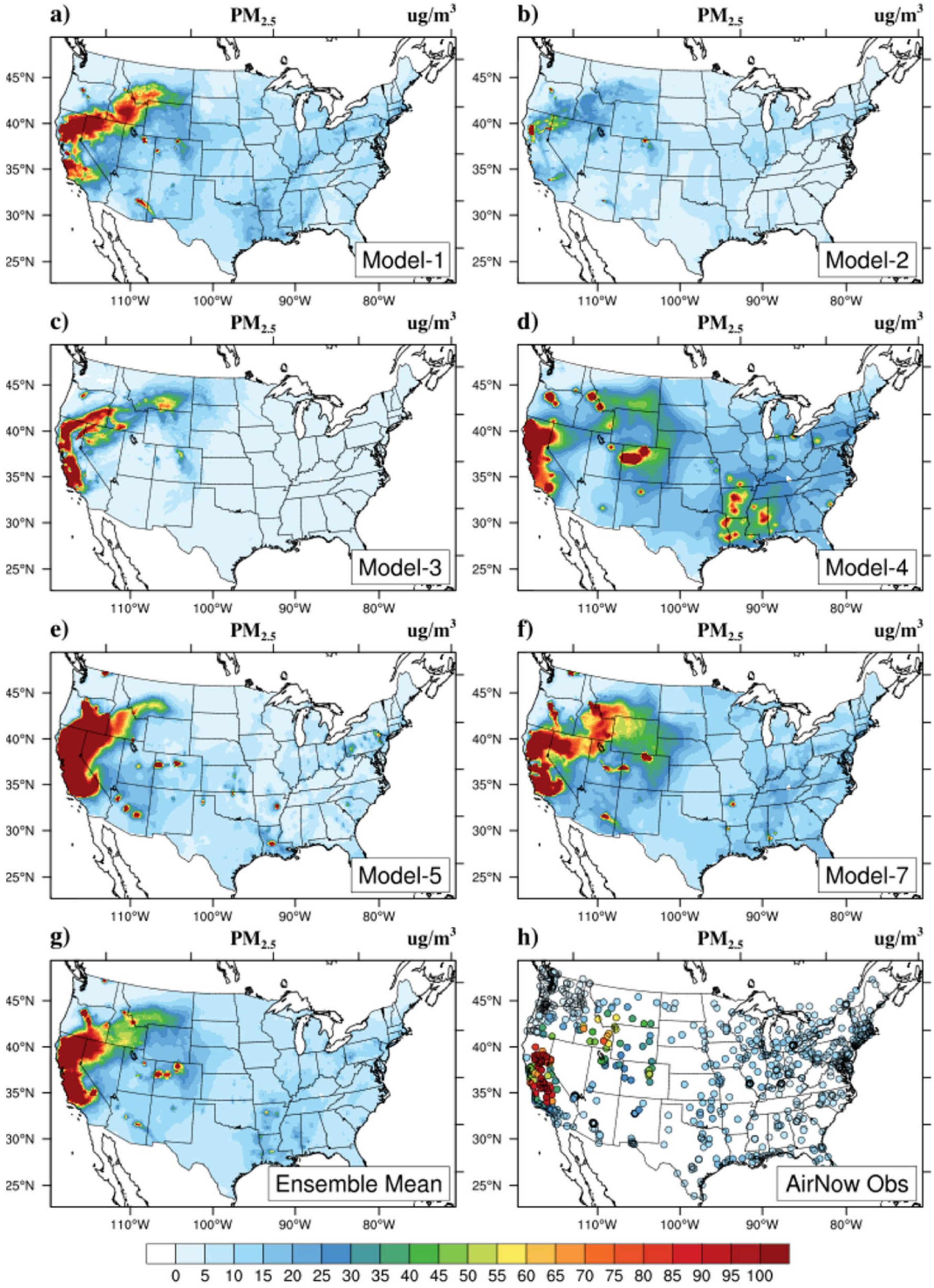

Figure 4. (a–f) Surface $PM_{2.5}$ concentrations predicted by six individual models and (g) the multi-model ensemble mean, compared with (h) AirNow $PM_{2.5}$ observations on 22 August 2020 (during the 2020 Gigafire events).

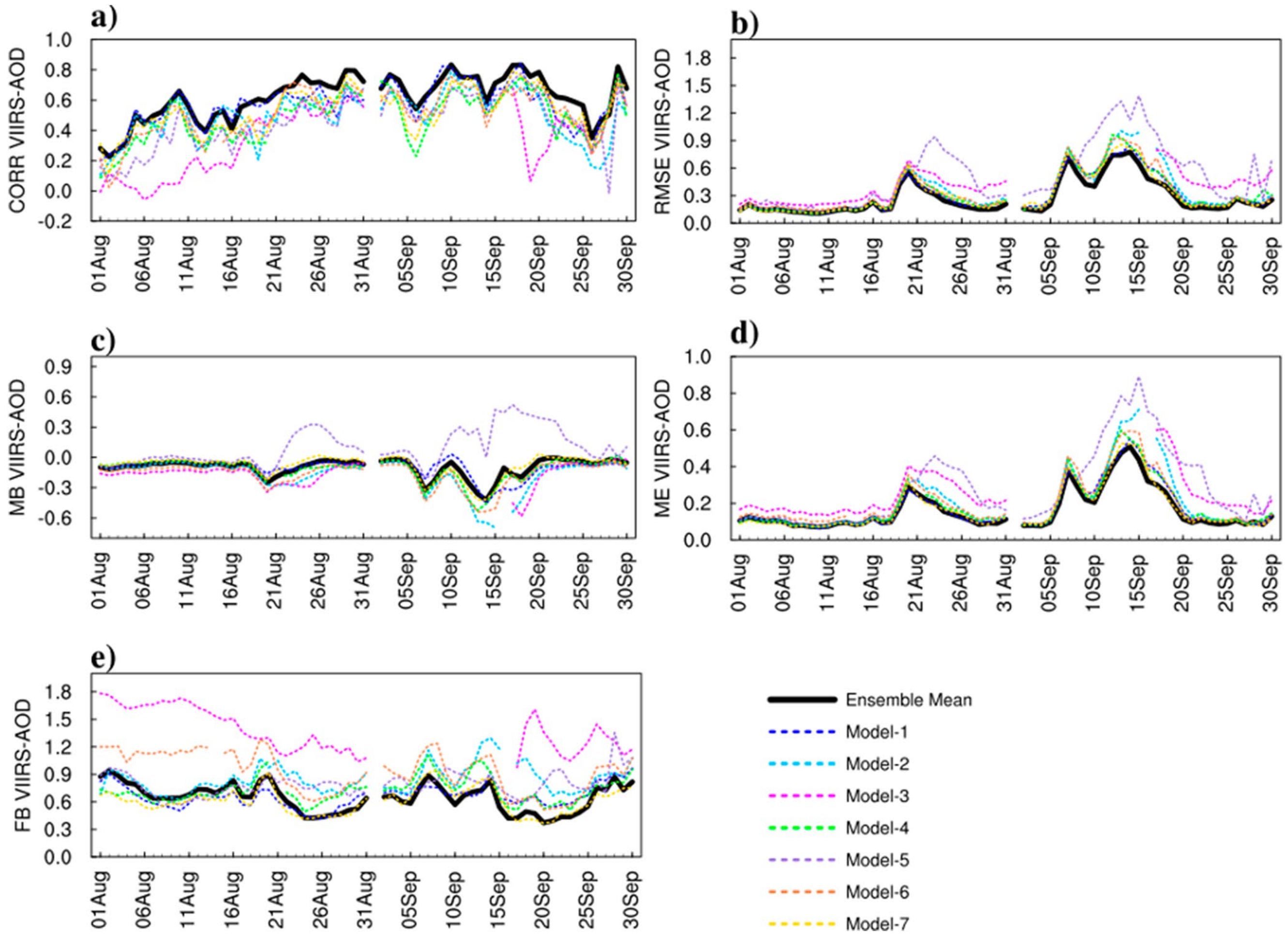

Figure 5. Time series of (a) root mean square error, (b) CORR, (c) mean bias (MB), (d) mean error (ME), and (e) fractional bias (FB) of aerosol optical depth (AOD) during the 2020 Gigafire events from August to September 2020. The AOD simulations by the ensemble mean (solid black line) and individual models (dash lines): Model-1 (blue), Model-2 (light blue), Model-3 (pink), Model-4 (green), Model-5 (purple), Model-6 (orange), and Model-7 (yellow) were compared against VIIRS AOD retrievals (Note: Model-3 shows a 6-week gap since its server had been down for 6 weeks).

higher correlation (CORR) during the intense wildfire period (mid-August and mid-September), and relatively lower root mean square error (RMSE; Figure 7b), mean bias (MB; Figure 7c), mean error (ME; Figure 7d), and fractional bias (FB; Figure 7g) for the entire period than most individual models (dashed lines). The positive MB values of the ensemble mean and the individual models in Figure 7c indicate the overestimation of $PM_{2.5}$ simulations for most of the time during the wildfire period, except for Model-2 and Model-3, which show negative MB values (underestimation of $PM_{2.5}$ concentrations). The errors in the simulations of $PM_{2.5}$ during the intense wildfire period were indicated by highly varying values of RMSE, ME, and FB in mid-August and mid-September, as shown in Figures 7b, 7d, and 7g.

Table 1 compares the eight statistical metrics calculated for surface $PM_{2.5}$ concentration by the individual models and the ensemble mean. Of the eight metrics, the ensemble mean scores the highest for four, including the two most important ones, the hit rate, and the overall rank. There is one member, Model-2, performing the best in terms of three of these metrics (RMSE, ME, and FAR). However, Model-2 has the lowest hit rate (accurate prediction of unhealthy air quality events), which makes its prediction the least useful for air quality warning during wildfire smoke events. Compared to individual models, the ensemble mean increased correlation and reduced bias and error in $PM_{2.5}$ forecasts, with the highest CORR (0.60) and scoring the best in FB (0.54), and lowering MB (7.40 μg/m^3), RMSE (25.93) and ME (11.16). Its overall rank is higher than any of these member models.

We further analyzed the performance of the MME to predict daily $PM_{2.5}$ exceedances (daily concentration >35 μg/m^3). As shown in Figures 7e and 7f, the ensemble mean (black plus signs) reduced aFAR (percentage

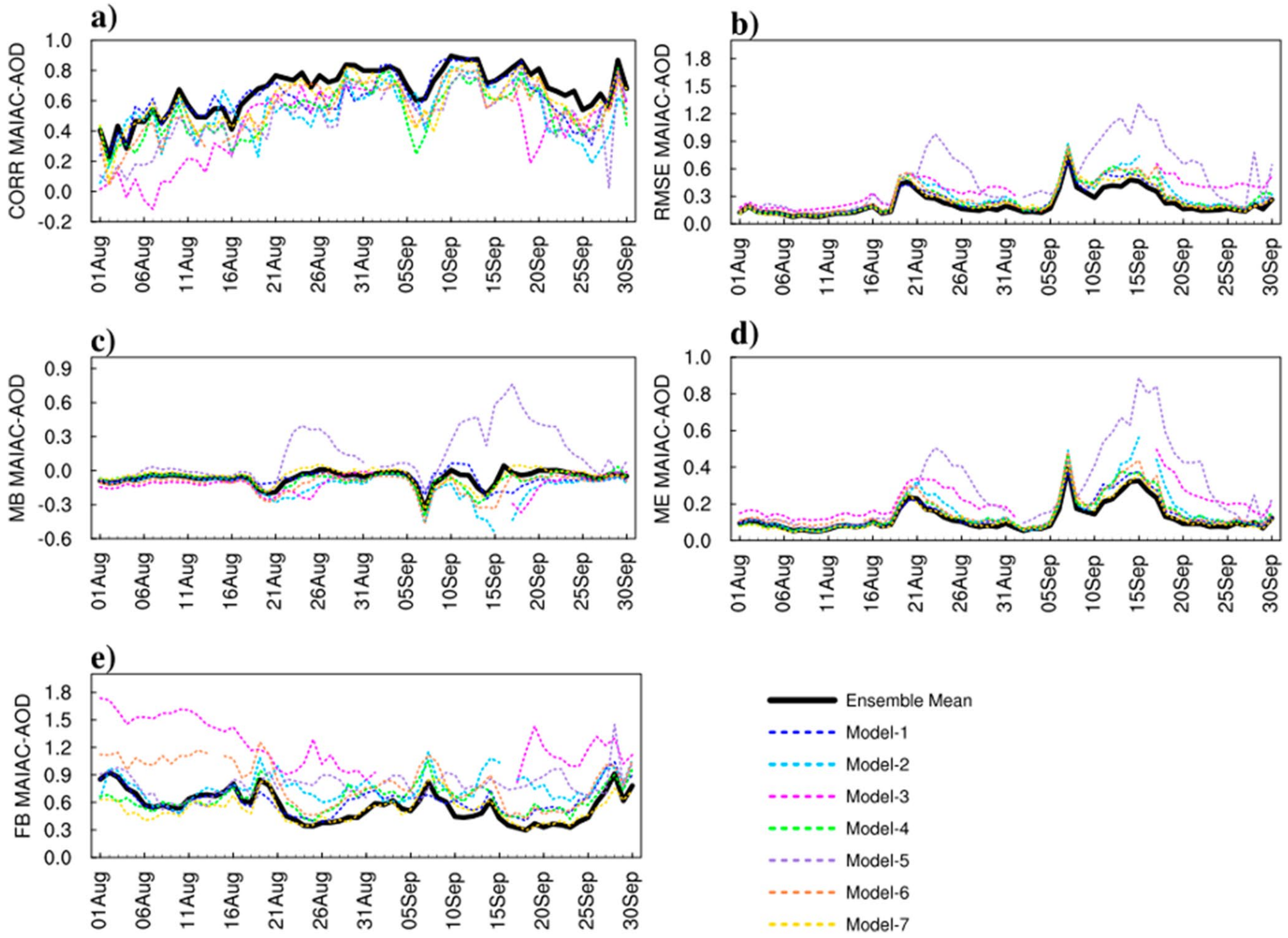

Figure 6. Time series of (a) root mean square error, (b) CORR, (c) mean bias (MB), (d) mean error (ME), and (e) fractional bias (FB) of aerosol optical depth (AOD) during the 2020 Gigafire events from August to September 2020. The AOD simulations by the ensemble mean (solid black line) and individual models (dash lines): Model-1 (blue), Model-2 (light blue), Model-3 (pink), Model-4 (green), Model-5 (purple), Model-6 (orange), and Model-7 (yellow) were compared against MAIAC AOD retrievals (Note: Model-3 shows a 6-week gap since its server had been down for 6 weeks).

of false alarms) and substantially increased the area hit rate (*aH*, percentage of hits), particularly in mid-August and mid-September when the extremely intense wildfires occurred, compared to individual models (colored plus signs). The ensemble mean achieves the highest *aH* value, predicting more than 86% of the observed $PM_{2.5}$ exceedances during extreme wildfires, and lowered *a*FAR to 60.52%. Due to relatively high correlation, high *aH*, low *a*FAR, and low FB values, the ensemble mean performs highly in RANK (2.83). These results suggest that the ensemble forecast has a practical advantage in reducing bias from individual forecasts of $PM_{2.5}$ and allowing effective probabilistic forecasts of $PM_{2.5}$. Furthermore, the evaluation results revealed that although a single model can be excellent at predicting AOD, it is not necessarily translated into good performance in surface $PM_{2.5}$ prediction. The model that performs highly in RANK for the AOD prediction is different from that of the $PM_{2.5}$ prediction.

3.3. Discussion of Ensemble Forecast Performance

Compared to individual models, the ensemble mean shows persistently higher RANK values for AOD and $PM_{2.5}$ throughout the study period (August–September 2020) when evaluated against three observation data sets: VIIRS (Figure 8a), MAIAC AOD (Figure 8b), and AirNow surface $PM_{2.5}$ concentrations (Figure 8c). The results suggest that the ensemble forecast is more reliable and performs better than individual model forecasts. In addition, it can reduce the bias (Table 1) because the ensemble mean cancels off positive and negative biases among individual model simulations. Note if most of the models underestimate (negative bias) or overestimate (positive bias)

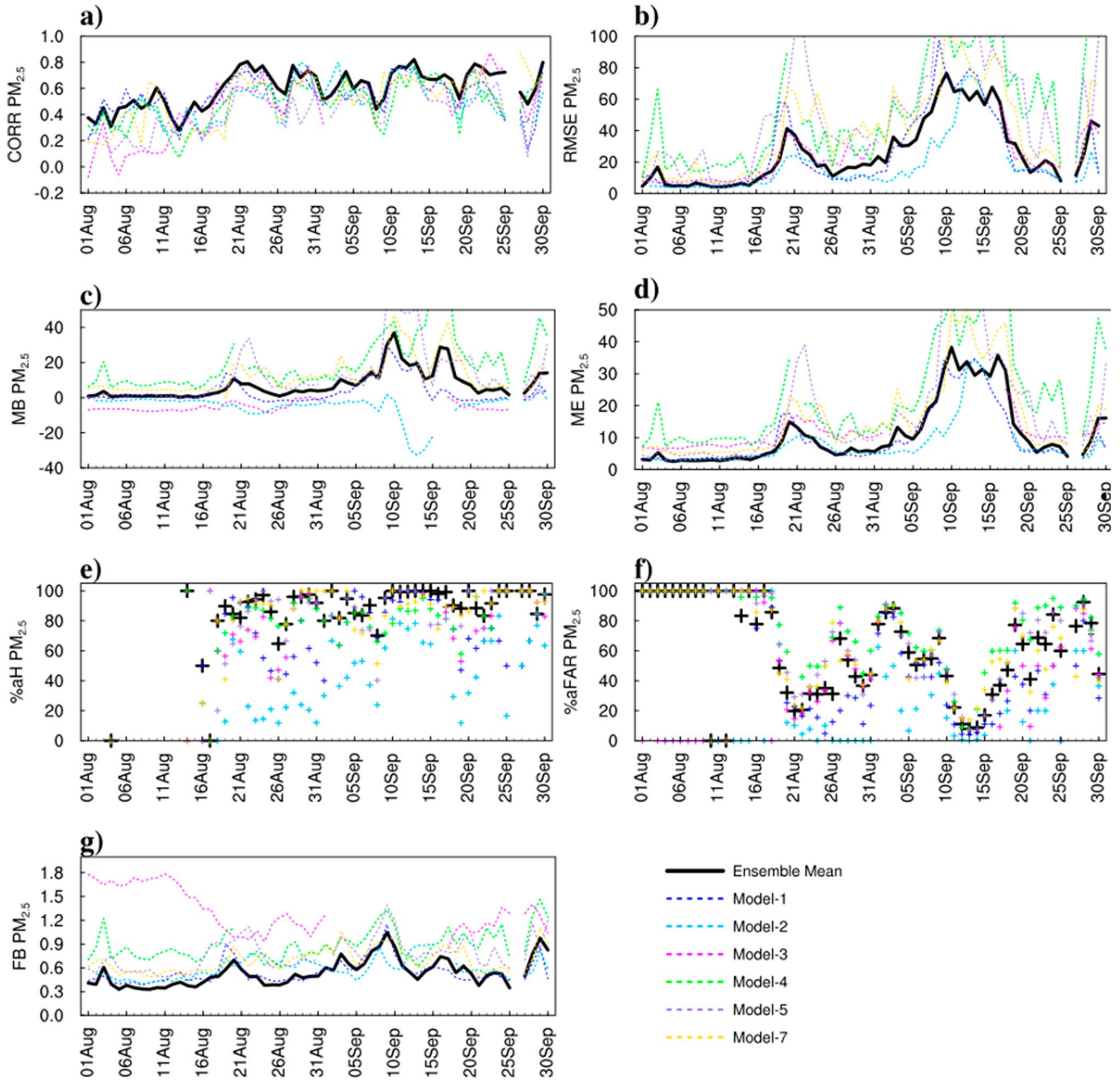

Figure 7. Time series of (a) root mean square error, (b) CORR, (c) mean bias (MB), (d) mean error (ME), (e) aH, (f) area false alarm ratio (aFAR), and (g) fractional bias (FB) of surface $PM_{2.5}$ concentration during the 2020 Gigafire events from August to September 2020. The $PM_{2.5}$ simulations by the ensemble mean (black solid line) and individual Model-1 (blue), Model-2 (light blue), Model-3 (pink), Model-4 (green), Model-5 (purple), and Model-7 (yellow) were compared against AirNow $PM_{2.5}$ observations (Note: Model-3 shows a 6-week gap since its server had been down for 6 weeks).

AOD and $PM_{2.5}$, the bias of the ensemble mean can become may worse than the top-ranked model. Therefore, the ensemble forecast is capable of improving forecasting only if there are complementary underestimations and overestimations by individual models.

For models that use satellite-based fire emission data sets, forest fire emissions are derived based on the latest satellite observations, which are assumed to continue during the forecasting period (24–120 hr, depending on the model). For real-time forecasting applications, the timeliness of the fire data is a key factor in determining the accuracy of wildfire air quality prediction (Hyer et al., 2023). Moreover, the intense wildfires during the middle of September 2020 generated very thick smoke that compromised the capability of satellite sensors to detect key

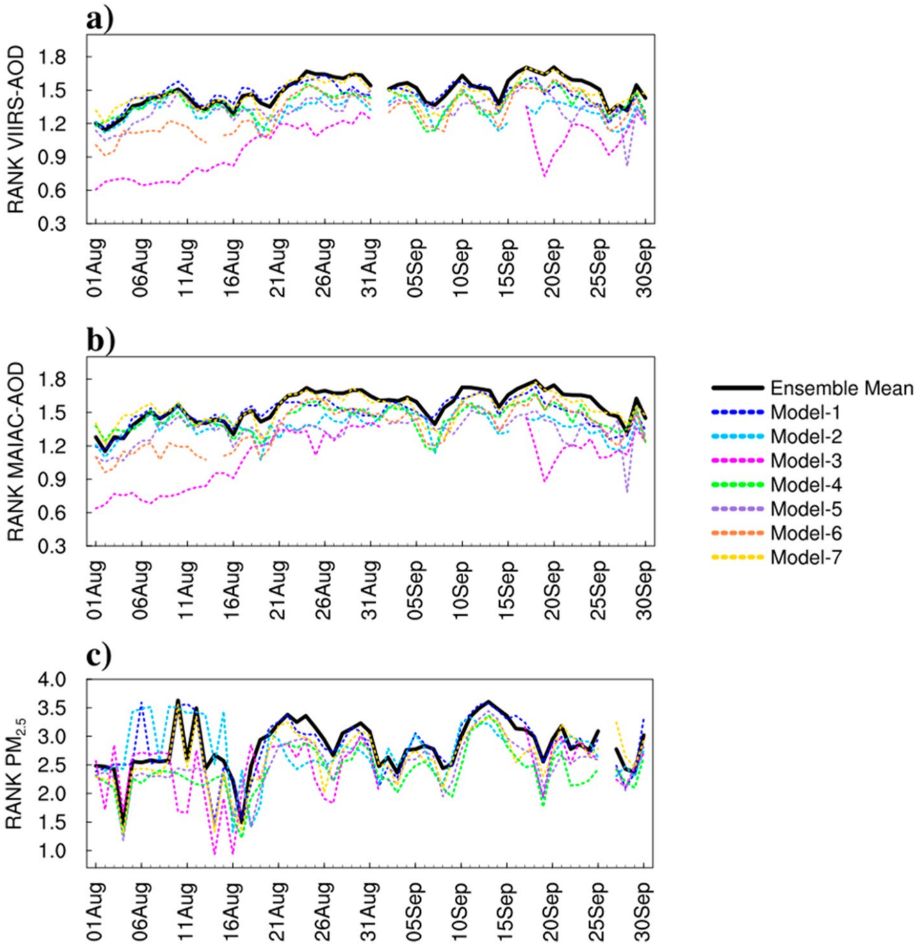

Figure 8. Time series of the overall rating (RANK) for aerosol optical depth (AOD) and $PM_{2.5}$ simulated by the ensemble mean and individual models. The RANK is calculated with four statistical metrics by comparing model predictions against AOD retrievals from (a) VIIRS and (b) MAIAC, and (c) surface $PM_{2.5}$ observations from AirNow during the 2020 Gigafire events from August to September 2020 (Note: Model-3 shows a 6-week gap since its server had been down for 6 weeks).

fire features such as fire hotspots, FRP, and AOD. These parameters are critical to either estimating fire emissions or evaluating model performance. The smoke could make the BB emissions applied to each model inaccurate and create a large error in smoke inventories (Kaiser et al., 2012). Furthermore, as the fires become stronger, the emissions are injected at a higher altitude, which often misrepresents vertical emissions within the PBL generated by each individual model (Ye et al., 2021). These two factors are important sources of uncertainties in air quality forecasts during wildfire events (Carter et al., 2020; X. Pan et al., 2020; Ye et al., 2021). Vernon et al. (2018) proposed that the plume injection height can affect smoke dispersion due to varying wind speeds and directions at different altitudes. Y. Li et al. (2020) found that a higher injection height can reduce near-source concentrations, and increase concentrations downwind. In addition, diurnal and day-to-day variations of wildfire behavior due to fuel aridity and availability, fire weather, fire containment activities and combustion stage can limit model forecasting performance during large wildfires (Saide et al., 2015). Besides, a variety of input data sets, such as meteorological fields and chemical transports (Garcia-Menendez et al., 2013; F. Li et al., 2019; Y. Li et al., 2020) and plume rise schemes (Briggs, 1969; Freitas et al., 2007; Y. Li et al., 2023; Paugam et al., 2016; Sofiev et al., 2012; Stein et al., 2009; Vernon et al., 2018; Zhu et al., 2018), are implemented differently in each model and can also impact the AOD and $PM_{2.5}$ forecasting performance (Delle Monache & Stull, 2003; Kumar et al., 2020).

Note the ensemble forecast calculates the mean of individual forecasts without applying any weighting factors. We also constructed a weighted ensemble. The weights for each individual model were determined by minimizing the differences between observations and model simulations using the multilinear regression method. When compared to the unweighted ensemble mean, the weighted ensemble exhibited mixed performance results (see

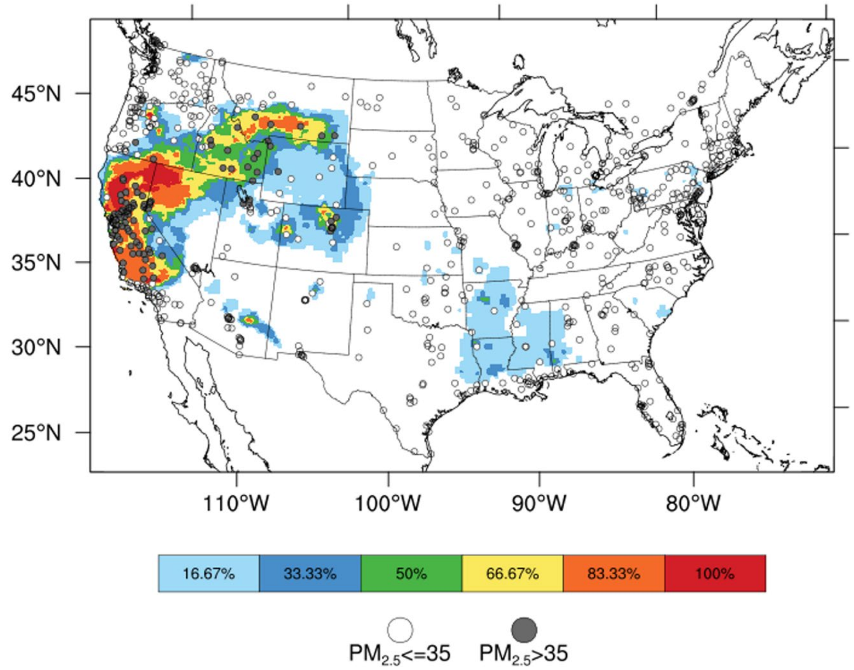

Figure 9. Ensemble probability forecast of $PM_{2.5}$ exceedances on 22 August 2020 (during the 2020 Gigafire events). Foreground colors indicate the probability values ranging from 16.67% (one out of six models forecasts the $PM_{2.5}$ exceedance or the exceedances are very unlikely to occur) (light blue) to 100% (all six models forecast the $PM_{2.5}$ exceedances or the exceedances are very likely to occur) (red). The $PM_{2.5}$ exceedances observed by the AirNow sites are displayed in gray/white circles (gray means an exceedance recorded by the monitor, and white means no exceedance recorded).

Table S2 in Supporting Information S1). While it effectively reduced biases, including RMSE, MB, and NMB, it did not improve performance for predicting extreme events, for example, the accuracy to predicting exceedances of the NAAQS for $PM_{2.5}$. Notably, the area hit ratio of the weighted ensemble was 20% lower than that of the unweighted ensemble. This discrepancy may arise from the fact that the weights calculated for the entire CONUS domain might not be suitable for specific regions. In addition, the Gigafire event studied here is unprecedented in many ways, including the burned areas, emission amounts, fire intensity, and burning duration. Because of the exceptionalness of this event, there is no historical training data available to derive better weighting factors, which further prevents us from switching to the weighted method. Future studies should explore the potential of region-specific weighting factors, as well as different methods to establish stable weights that can be reliably used for real-time forecasting.

3.4. Ensemble Probability Forecast of $PM_{2.5}$ Exceedances

In general, the ensemble probability (Equation 1) shows fairly good performance in forecasting $PM_{2.5}$ exceedances during the 2020 Gigafire events. Figure 9 depicts a contour map of ensemble probability forecast values overlaid by the actual exceedance (binary) over the AirNow sites across the CONUS. The exceedance probability ranges from 0% (no exceedances predicted by any models) to 100% (exceedances predicted by all six models). The larger the number of models that forecast the exceedance for each grid, the higher probability that the exceedances will occur in that grid. As shown in Figure 9, the contours of high ensemble probability values of 83.33% (exceedances are likely to occur; orange) and 100% (exceedances are very likely to occur; red) were displayed mainly in California, which collocated well with the AirNow exceedance measurements (marked as filled gray circles). However, the AirNow observed exceedances in the downwind region (Idaho and Montana) were only captured by four of the six models, giving a probability forecast of 66.67% (exceedances probably occur; yellow). The degradation of exceedance probability in the downwind areas highlights the challenges in predicting transported smoke plumes and their effects on surface air quality.

We also evaluated the performance of forecasting $PM_{2.5}$ exceedances during extreme fire events by comparing the predicted ensemble exceedance probability against the AirNow observed $PM_{2.5}$ exceedances. The results are

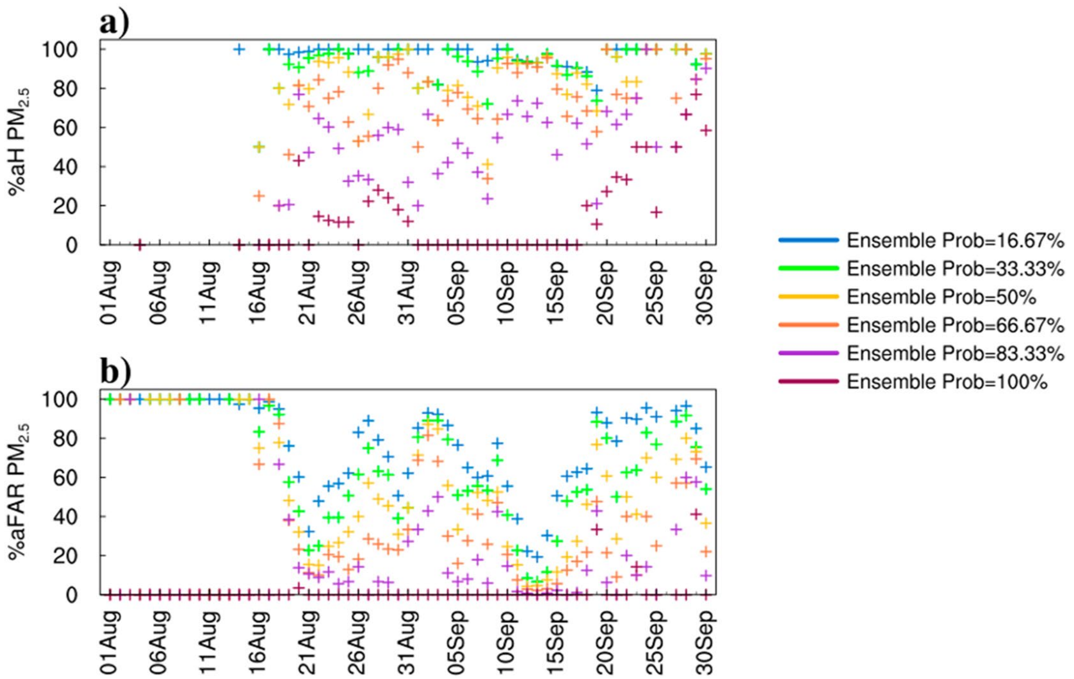

Figure 10. Time series plots of (a) *aH* (percentage of hits) and (b) area false alarm ratio (*a*FAR) (Percentage of false alarms) during the 2020 Gigafire events (August–September 2020) for the ensemble probability forecast of $PM_{2.5}$ exceedances. Ensemble probability values range from 16.67% (one out of six models or the exceedances are very unlikely to occur) to 100% (all six models or the exceedances are very likely to occur).

shown as time series plots of *aH* and *a*FAR in Figures 10a and 10b. The average *aH* (percentage of hits) and *a*FAR (percentage of false alarms) values are listed in Table 2. High *aH* and low *a*FAR suggest good agreement between model simulated exceedances and observed exceedances. As displayed in the time series plots of *aH* and *a*FAR (Figures 10a and 10b) and the average *aH* and *a*FAR values (Table 2), the lowest ensemble probability of 16.67% (exceedances are very unlikely to occur; blue plus sign) is constantly associated with high *aH* and high *a*FAR throughout the study period, resulting in being top-ranked *aH* (93.99%) and the lowest-ranked *a*FAR (78.00%) on average, while the highest ensemble probability of 100% (exceedances are very likely to occur) show persistently and relatively low *aH* and low *a*FAR all the time, resulting in holding the lowest-ranked *aH* (14.73%) and the top-ranked *a*FAR (1.54%).

Table 2
Averaged aH (Hit Rate) and aFAR (False Alarm Rate) of Ensemble Probability Forecast of $PM_{2.5}$ Exceedances Forecast During the 2020 Gigafire Events (August–September 2020) Calculated From Comparing Simulated $PM_{2.5}$ Exceedances and Observed $PM_{2.5}$ Exceedances Obtained From AirNow

	Statistical metric	
Ensemble probability	aH (%)	aFAR (%)
16.67% (1 of 6 models)	93.99	78.00
33.33% (2 of 6 models)	88.40	63.33
50% (3 of 6 models)	79.31	47.51
66.67% (4 of 6 models)	69.72	28.99
83.33% (5 of 6 models)	48.10	15.38
100% (all models)	14.73	1.54

Note. The probability value is based on the number of ensemble members that predict the exceedances (Note that the total number of ensemble members is six).

The evaluation results imply that the low ensemble probability shows better performance in forecasting observed exceedances across the CONUS. This is because some exceedances predicted by a small subset of models were true exceedances associated with wildfires, especially in the wildfire active regions, resulting in high *aH* (percentage of hits). However, the remaining exceedances predicted elsewhere were false alarms influenced by overestimation that could not be removed from the forecast due to a lack of calibration and validation with other models. As a result, the lowest probability values generally yield high *a*FAR (percentage of false alarms). Conversely, the ensemble forecast with a larger number of models or the higher ensemble probability performs more accurately and reliably in forecasting $PM_{2.5}$ exceedances on a smaller or local scale because their predicted exceedances have been calibrated and verified with the co-existed exceedances predicted by the other participating models included in the ensemble. As a consequence, the areas showing false exceedances have been reduced or removed, resulting in lower *aH* and *a*FAR.

Practically, the accuracy of the exceedance probability forecasts depends on the original grid resolution of each ensemble member. The exceedances simulated by the global models generally cover larger areas compared to that by the regional models, even after being interpolated to a higher spatial resolution. Using the MME approach can satisfactorily reduce the biases in the exceedance probability forecast, which are frequently a result of the discrepancies between the spatial resolutions of the ensemble members. Furthermore, the statistics results also highlight that the ensemble exceedance probability forecast has the potential to provide complementarily estimated high-risk areas (the areas within the ensemble exceedance probability of 100%) associated with $PM_{2.5}$ exceedances during the wildfires in addition to the ground observations, especially in the location where the monitoring sites are sparse or nonexistent.

4. Conclusions

Wildfires are important natural emission sources that contribute large amounts of aerosols to the atmosphere that exerts detrimental impacts on society, such as adverse health effects, life and property losses, and disruption of economic activities. In this study, we developed and evaluated a North America ensemble wildfire smoke forecasting system to improve the predictability of wildfire AOD and $PM_{2.5}$. The MME forecasts were built using three regional models, one global ensemble model, and three global models operated by NASA, NOAA, NRL, and GMU. Our ensemble forecast reproduces daily forecasts of AOD and $PM_{2.5}$ as well as the probability of $PM_{2.5}$ exceedances (daily concentration >35 μg/m^3) on a 12 × 12 km grid resolution over the Contiguous U.S. (CONUS) during the 2020 Gigafire events (August–September 2020) in the western U.S.

The performance of the ensemble forecast for AOD and $PM_{2.5}$ was evaluated with AOD retrievals from the VIIRS and MODIS-MAIAC and $PM_{2.5}$ measurements from the AirNow network. A suite of statistical metrics, including five single metrics (RMSE, CORR, MB, ME, and FB), an overall rating (RANK), and two discrete categorical metrics (area hit rate; *aH* and aFAR) were employed to measure the performance of ensemble mean and ensemble probability in predicting the exceedances of the NAAQS for $PM_{2.5}$ during the 2020 Gigafire events. The results suggested that the ensemble mean, compared to the individual forecasts, can significantly reduce the biases and uncertainties in the wildfire air pollution forecast and produce more reliable forecasts during the study period. For AOD forecasts, the ensemble mean was able to improve model performance, as indicated by reduced bias and error, and the strongest correlation. The ensemble mean also achieved the best overall RANK when compared against VIIRS (1.48 from a range of 0.98–1.46 by individual forecasts) and MAIAC AOD (1.55 from a range of 1.08–1.53 by individual forecasts). For surface $PM_{2.5}$, the ensemble mean outperformed all individual models, with the strongest correlation (0.60 vs. 0.43–0.54 by individual forecasts), the lowest FB (0.54 vs. 0.55–1.32), the highest area hit rate (87% vs. 40%–82%), decreased MB (7.40 μg/m^3) and the best overall RANK (2.83 from a range of 2.40–2.81). In terms of the exceedance probability forecasting (binary prediction) performance, the ensemble practically generated a well suited exceedance probability forecast that matched the observed AirNow exceedances fairly well, as demonstrated by the lowest false alarm rate (*a*FAR) at 1.52% achieved by the ensemble probability of 100%. This result suggested a great potential of the ensemble exceedance probability forecast to provide air pollution early warning alerts when the $PM_{2.5}$ concentrations exceed the health-based NAAQS (daily concentration >35 μg/m^3) during wildfire events.

The results of this paper demonstrate that even with the relatively straightforward MME mean method, which has been widely used in the regional and global ensemble of atmospheric composition predictions (e.g., Sessions et al., 2015; Xian et al., 2019), the ensemble approach can yield better prediction during the unprecedented Gigafire events. Such a finding is practically useful, without demanding too many extra resources, for the federal agencies to provide better early warning services to the public.

Note that the MME wildfire air quality forecast system presented here is still at an early stage for real-time deployment over North America. It is necessary to extend the ensemble forecast to other fires and periods, including the 2021 fire season and more recent Canadian wildfire events, to examine if this method can be applied in other cases. The intercomparison between the ensemble and individual models can be useful to investigate differences in emissions, meteorology inputs, and plume rise algorithms, as well as chemical transport/dispersion model processes among these models. In addition, we will explore various methods to develop the ensemble forecasts, such as exploring weighted ensemble and comparing weighted and unweighted ensemble means. Collectively, these research efforts will lead to improved real-time wildfire smoke forecasts which will be over North America to support key decision-making on air quality and public health at local, national, and international levels.

Data Availability Statement

The data for VIIRS, GEOS5, HYSPLIT, NAAPS, NACC-CMAQ, and GMU-CMAQ can be downloaded from https://zenodo.org/records/10126149.

The AirNow observations can be downloaded from: https://files.airnowtech.org/?prefix=airnow/2020/.

The GEFS-Aerosols results can be downloaded from: https://noaa-gefs-pds.s3.amazonaws.com/index.html.

The ICAP-MME results can be downloaded from: https://nrlgodae1.nrlmry.navy.mil/ftp/outgoing/nrl/ICAP-MME/.

Acknowledgments

This study is financially supported by NASA Health and Air Quality Program (80NSSC21K0512), NOAA Weather Program Office, and Disaster Supplemental Program (NA20OAR4600275). We thank NASA, NOAA, and NRL for providing the model prediction data used for constructing the ensemble forecast. Ground measurements collected by the EPA and satellite AOD data by NOAA and NASA are gratefully acknowledged. We are grateful to three anonymous reviewers to provide constructive comments on earlier versions of this manuscript. The views expressed are those of the authors and are not reflective of the federal agencies (NOAA, NASA, NRL, etc.) or institutions.

References

Benedetti, A., Reid, J. S., & Colarco, P. R. (2011). International cooperative for aerosol prediction workshop on aerosol forecast verification. *Bulletin of the American Meteorological Society*, *92*(11), ES48–ES53. https://doi.org/10.1175/BAMS-D-11-00105.1

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

Buchard, V., Randles, C. A., da Silva, A. M., Darmenov, A., Colarco, P. R., Govindaraju, R., et al. (2017). The MERRA-2 aerosol reanalysis, 1980 onward. Part II: Evaluation and case studies. *Journal of Climate*, *30*(17), 6851–6872. https://doi.org/10.1175/JCLI-D-16-0613.1

California Department of Forestry and Fire Protection (CAL FIRE). (2020). 2020 fire season. Retrieved from https://www.fire.ca.gov/incidents/2020/

Campbell, P. C., Tang, Y., Lee, P., Baker, B., Tong, D., Saylor, R., et al. (2022). Development and evaluation of an advanced National Air Quality Forecasting Capability using the NOAA Global Forecast System version 16. *Geoscientific Model Development*, *15*(8), 3281–3313. https://doi.org/10.5194/gmd-15-3281-2022

Cao, C., De Luccia, F. J., Xiong, X., Wolfe, R., & Weng, F. (2014). Early on-orbit performance of the visible infrared imaging radiometer suite onboard the Suomi National Polar-Orbiting Partnership (S-NPP) satellite. *IEEE Transactions on Geoscience and Remote Sensing*, *52*(2), 1142–1156. https://doi.org/10.1109/TGRS.2013.2247768

Cao, C., Xiong, J., Blonski, S., Liu, Q., Uprety, S., Shao, X., et al. (2013). Suomi NPP VIIRS sensor data record verification, validation, and long-term performance monitoring. *Journal of Geophysical Research: Atmospheres*, *118*(20), 11664–11678. https://doi.org/10.1002/2013JD020418

Carter, T. S., Heald, C. L., Jimenez, J. L., Campuzano-Jost, P., Kondo, Y., Moteki, N., et al. (2020). How emissions uncertainty influences the distribution and radiative impacts of smoke from fires in North America. *Atmospheric Chemistry and Physics*, *20*(4), 2073–2097. https://doi.org/10.5194/acp-20-2073-2020

Chin, M., Ginoux, P., Kinne, S., Torres, O., Holben, B. N., Duncan, B. N., et al. (2002). Tropospheric aerosol optical thickness from the GOCART model and comparisons with satellite and sun photometer measurements. *Journal of the Atmospheric Sciences*, *59*(3), 461–483. https://doi.org/10.1175/1520-0469(2002)059<0461:TAOTFT>2.0.CO;2

Colarco, P., Benedetti, A., Reid, J., & Tanaka, T. (2014). Using EOS data to improve aerosol forecasting: The International Cooperative for Aerosol Research (ICAP). *The Earth Observer*, *26*, 14–19.

Colarco, P., da Silva, A., Chin, M., & Diehl, T. (2010). Online simulations of global aerosol distributions in the NASA GEOS-4 model and comparisons to satellite and ground-based aerosol optical depth. *Journal of Geophysical Research*, *115*(D14), D14207. https://doi.org/10.1029/2009JD012820

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

Delle Monache, L., Alessandrini, S., Djalalova, I., Wilczak, J., Knievel, J. C., & Kumar, R. (2020). Improving air quality predictions over the United States with an analog ensemble. *Weather and Forecasting*, *35*(5), 2145–2162. https://doi.org/10.1175/WAF-D-19-0148.1

Delle Monache, L., Deng, X., Zhou, Y., & Stull, R. (2006). Ozone ensemble forecasts: 1. A new ensemble design. *Journal of Geophysical Research*, *111*(D5), D05307. https://doi.org/10.1029/2005JD006310

Delle Monache, L., Nipen, T., Deng, X., Zhou, Y., & Stull, R. (2006). Ozone ensemble forecasts: 2. A Kalman filter predictor bias correction. *Journal of Geophysical Research*, *111*(D5), D05308. https://doi.org/10.1029/2005JD006311

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

Delle Monache, L., Wilczak, J., Mckeen, S., Grell, G., Pagowski, M., Peckham, S., et al. (2008). A Kalman-filter bias correction method applied to deterministic, ensemble averaged and probabilistic forecasts of surface ozone. *Tellus B: Chemical and Physical Meteorology*, *60*(2), 238–249. https://doi.org/10.1111/j.1600-0889.2007.00332.x

Draxler, R. R. (2006). The use of global and mesoscale meteorological model data to predict the transport and dispersion of tracer plumes over Washington, D.C. *Weather and Forecasting*, *21*(3), 383–394. https://doi.org/10.1175/WAF926.1

Eyth, A., Vukovich, J., Farkas, C., & Strum, M. (2020). Technical Support Document (TSD) preparation of emissions inventories for 2016v1 North American emissions modeling platform.

Fann, N., Alman, B., Broome, R. A., Morgan, G. G., Johnston, F. H., Pouliot, G., & Rappold, A. G. (2018). The health impacts and economic value of wildland fire episodes in the U.S.: 2008–2012. *Science of the Total Environment*, *610–611*, 802–809. https://doi.org/10.1016/j.scitotenv.2017.08.024

Ford, B., Martin, M. V., Zelasky, S. E., Fischer, E. V., Anenberg, S. C., Heald, C. L., & Pierce, J. R. (2018). Future fire impacts on smoke concentrations, visibility, and health in the Contiguous United States. *GeoHealth*, *2*(8), 229–247. https://doi.org/10.1029/2018GH000144

Freitas, S. R., Longo, K. M., Chatfield, R., Latham, D., Silva Dias, M. A. F., Andreae, M. O., et al. (2007). Including the sub-grid scale plume rise of vegetation fires in low resolution atmospheric transport models. *Atmospheric Chemistry and Physics*, *7*(13), 3385–3398. https://doi.org/10.5194/acp-7-3385-2007

Garcia-Menendez, F., Hu, Y., & Odman, M. T. (2013). Simulating smoke transport from wildland fires with a regional-scale air quality model: Sensitivity to uncertain wind fields. *Journal of Geophysical Research: Atmospheres*, *118*(12), 6493–6504. https://doi.org/10.1002/jgrd.50524

Goerss, J., Sampson, C., & Gross, J. (2004). A history of western North Pacific tropical cyclone track forecast skill. *Weather and Forecasting*, *19*(3), 633–638. https://doi.org/10.1175/1520-0434(2004)019<0633:AHOWNP>2.0.CO;2

Guirguis, K., Gershunov, A., Cayan, D. R., & Pierce, D. W. (2018). Heat wave probability in the changing climate of the Southwest US. *Climate Dynamics*, *50*(9–10), 3853–3864. https://doi.org/10.1007/s00382-017-3850-3

Hamill, T. M., Whitaker, J. S., Fiorino, M., & Benjamin, S. G. (2011). Global ensemble predictions of 2009's Tropical cyclones initialized with an ensemble Kalman Filter. *Monthly Weather Review*, *139*(2), 668–688. https://doi.org/10.1175/2010MWR3456.1

Hamill, T. M., Whitaker, J. S., Kleist, D. T., Fiorino, M., & Benjamin, S. G. (2011). Predictions of 2010's tropical cyclones using the GFS and ensemble-based data assimilation methods. *Monthly Weather Review*, *139*(10), 3243–3247. https://doi.org/10.1175/MWR-D-11-00079.1

Hessburg, P. F., Miller, C. L., Parks, S. A., Povak, N. A., Taylor, A. H., Higuera, P. E., et al. (2019). Climate, environment, and disturbance history govern resilience of western North American forests. *Frontiers in Ecology and Evolution*, *7*, 239. https://doi.org/10.3389/fevo.2019.00239

Hogan, T., Liu, M., Ridout, J., Peng, M., Whitcomb, T., Ruston, B., et al. (2014). The navy global environmental model. *Oceanography*, *27*(3), 116–125. https://doi.org/10.5670/oceanog.2014.73

Houyoux, M. R., Vukovich, J. M., Coats, C. J., Wheeler, N. J. M., & Kasibhatla, P. S. (2000). Emission inventory development and processing for the Seasonal Model for Regional Air Quality (SMRAQ) project. *Journal of Geophysical Research*, *105*(D7), 9079–9090. https://doi.org/10.1029/1999JD900975

Hulley, G. C., Dousset, B., & Kahn, B. H. (2020). Rising trends in heatwave metrics across Southern California. *Earth's Future*, *8*(7), e2020EF001480. https://doi.org/10.1029/2020EF001480

Hyer, E. J., Camacho, C. P., Peterson, D. A., Satterfield, E. A., & Saide, P. E. (2023). Data assimilation for numerical smoke prediction. In T. V. Loboda, N. H. F. French, & R. C. Puett (Eds.), *Landscape fire, smoke, and health*. https://doi.org/10.1002/9781119757030.ch7

Hyer, E. J., Reid, J. S., & Zhang, J. (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

Kaiser, J. W., Heil, A., Andreae, M. O., Benedetti, A., Chubarova, N., Jones, L., et al. (2012). Biomass burning emissions estimated with a global fire assimilation system based on observed fire radiative power. *Biogeosciences*, *9*(1), 527–554. https://doi.org/10.5194/bg-9-527-2012

Kang, D., Mathur, R., Schere, K., Yu, S., & Eder, B. (2007). New categorical metrics for air quality model evaluation. *Journal of Applied Meteorology and Climatology*, *46*(4), 549–555. https://doi.org/10.1175/JAM2479.1

Kumar, R., Alessandrini, S., Hodzic, A., & Lee, J. A. (2020). A novel ensemble design for probabilistic predictions of fine particulate matter over the Contiguous United States (CONUS). *Journal of Geophysical Research: Atmospheres*, *125*(16), e2020JD032554. https://doi.org/10.1029/2020JD032554

Larkin, N. K., O'Neill, S. M., Solomon, R., Raffuse, S., Strand, T., Sullivan, D. C., et al. (2009). The BlueSky smoke modeling framework. *International Journal of Wildland Fire*, *18*(8), 906. https://doi.org/10.1071/WF07086

Lee, P., McQueen, J., Stajner, I., Huang, J., Pan, L., Tong, D., et al. (2017). NAQFC developmental forecast guidance for fine particulate matter ($PM_{2.5}$). *Weather and Forecasting*, *32*(1), 343–360. https://doi.org/10.1175/WAF-D-15-0163.1

Levy, R. C., Mattoo, S., Munchak, L. A., Remer, L. A., Sayer, A. M., Patadia, F., & Hsu, N. C. (2013). The collection 6 MODIS aerosol products over land and ocean. *Atmospheric Measurement Techniques*, *6*(11), 2989–3034. https://doi.org/10.5194/amt-6-2989-2013

Levy, R. C., Munchak, L. A., Mattoo, S., Patadia, F., Remer, L. A., & Holz, R. E. (2015). Towards a long-term global aerosol optical depth record: Applying a consistent aerosol retrieval algorithm to MODIS and VIIRS-observed reflectance. *Atmospheric Measurement Techniques*, *8*(10), 4083–4110. https://doi.org/10.5194/amt-8-4083-2015

Li, F., Val Martin, M., Andreae, M. O., Arneth, A., Hantson, S., Kaiser, J. W., et al. (2019). Historical (1700–2012) global multi-model estimates of the fire emissions from the Fire Modeling Intercomparison Project (FireMIP). *Atmospheric Chemistry and Physics*, *19*(19), 12545–12567. https://doi.org/10.5194/acp-19-12545-2019

Li, S., & Banerjee, T. (2021). Spatial and temporal pattern of wildfires in California from 2000 to 2019. *Scientific Reports*, *11*(1), 8779. https://doi.org/10.1038/s41598-021-88131-9

Li, Y., Tong, D., Ma, S., Freitas, S. R., Ahmadov, R., Sofiev, M., et al. (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). *Atmospheric Chemistry and Physics*, *23*(5), 3083–3101. https://doi.org/10.5194/acp-23-3083-2023

Li, Y., Tong, D., Ma, S., Zhang, X., Kondragunta, S., Li, F., & Saylor, R. (2021). Dominance of wildfires impact on air quality exceedances during the 2020 record-breaking wildfire season in the United States. *Geophysical Research Letters*, *48*(21), e2021GL094908. https://doi.org/10.1029/2021GL094908

Li, Y., Tong, D. Q., Ngan, F., Cohen, M. D., Stein, A. F., Kondragunta, S., et al. (2020). Ensemble $PM_{2.5}$ forecasting during the 2018 camp fire event using the HYSPLIT transport and dispersion model. *Journal of Geophysical Research: Atmospheres*, *125*(15), e2020JD032768. https://doi.org/10.1029/2020JD032768

Liu, Y., Goodrick, S. L., & Stanturf, J. A. (2013). Future U.S. wildfire potential trends projected using a dynamically downscaled climate change scenario. *Forest Ecology and Management*, *294*, 120–135. https://doi.org/10.1016/j.foreco.2012.06.049

Lyapustin, A., Korkin, S., Wang, Y., Quayle, B., & Laszlo, I. (2012). Discrimination of biomass burning smoke and clouds in MAIAC algorithm. *Atmospheric Chemistry and Physics*, *12*(20), 9679–9686. https://doi.org/10.5194/acp-12-9679-2012

Lyapustin, A., Martonchik, J., Wang, Y., Laszlo, I., & Korkin, S. (2011). Multiangle Implementation of Atmospheric Correction (MAIAC): 1. Radiative transfer basis and look-up tables. *Journal of Geophysical Research*, *116*(D3), D03210. https://doi.org/10.1029/2010JD014985

Lyapustin, A., Wang, Y., Korkin, S., & Huang, D. (2018). MODIS collection 6 MAIAC algorithm. *Atmospheric Measurement Techniques*, *11*(10), 5741–5765. https://doi.org/10.5194/amt-11-5741-2018

Lyapustin, A., Wang, Y., Korkin, S., Kahn, R., & Winker, D. (2020). MAIAC thermal technique for smoke injection height from MODIS. *IEEE Geoscience and Remote Sensing Letters*, *17*(5), 730–734. https://doi.org/10.1109/LGRS.2019.2936332

Lyapustin, A., Wang, Y., Laszlo, I., Kahn, R., Korkin, S., Remer, L., et al. (2011). Multiangle Implementation of Atmospheric Correction (MAIAC): 2. Aerosol algorithm. *Journal of Geophysical Research*, *116*(D3), D03211. https://doi.org/10.1029/2010JD014986

Lyapustin, A., Wang, Y., Xiong, X., Meister, G., Platnick, S., Levy, R., et al. (2014). Scientific impact of MODIS C5 calibration degradation and C6+ improvements. *Atmospheric Measurement Techniques*, *7*(12), 4353–4365. https://doi.org/10.5194/amt-7-4353-2014

Lynch, P., Reid, J. S., Westphal, D. L., Zhang, J., Hogan, T. F., Hyer, E. J., et al. (2016). An 11-year global gridded aerosol optical thickness reanalysis (v1.0) for atmospheric and climate sciences. *Geoscientific Model Development*, *9*(4), 1489–1522. https://doi.org/10.5194/gmd-9-1489-2016

McClure, C. D., & Jaffe, D. A. (2018). US particulate matter air quality improves except in wildfire-prone areas. *Proceedings of the National Academy of Sciences*, *115*(31), 7901–7906. https://doi.org/10.1073/pnas.1804353115

National Interagency Coordination Center (NICC). (2020). Wildland fire summary and statistics annual reports 2020. Retrieved from https://www.predictiveservices.nifc.gov/intelligence/2020_statssumm/

Neumann, J. E., Amend, M., Anenberg, S., Kinney, P. L., Sarofim, M., Martinich, J., et al. (2021). Estimating $PM_{2.5}$-related premature mortality and morbidity associated with future wildfire emissions in the western US. *Environmental Research Letters*, *16*(3), 035019. https://doi.org/10.1088/1748-9326/abe82b

NIFC National Interagency Coordination Center. (2020). Wildland fire summary and statistics annual report 2020. Retrieved from https://www.predictiveservices.nifc.gov/intelligence/2020_statssumm/annual_report_2020.pdf

Pan, L., Tong, D., Lee, P., Kim, H.-C., & Chai, T. (2014). Assessment of NOx and O_3 forecasting performances in the U.S. National Air Quality Forecasting Capability before and after the 2012 major emissions updates. *Atmospheric Environment*, *95*, 610–619. https://doi.org/10.1016/j.atmosenv.2014.06.020

Pan, X., Ichoku, C., Chin, M., Bian, H., Darmenov, A., Colarco, P., et al. (2020). Six global biomass burning emission datasets: Intercomparison and application in one global aerosol model. *Atmospheric Chemistry and Physics*, *20*(2), 969–994. https://doi.org/10.5194/acp-20-969-2020

Pathak, T. B., Maskey, M. L., Dahlberg, J. A., Kearns, F., Bali, K. M., & Zaccaria, D. (2018). Climate change trends and impacts on California agriculture: A detailed review. *Agronomy*, *8*(3), 25. https://doi.org/10.3390/agronomy8030025

Paugam, R., Wooster, M., Freitas, S., & Val Martin, M. (2016). A review of approaches to estimate wildfire plume injection height within large-scale atmospheric chemical transport models. *Atmospheric Chemistry and Physics*, *16*(2), 907–925. https://doi.org/10.5194/acp-16-907-2016

Petersen, A. K., Brasseur, G. P., Bouarar, I., Flemming, J., Gauss, M., Jiang, F., et al. (2019). Ensemble forecasts of air quality in eastern China – Part 2: Evaluation of the MarcoPolo–Panda prediction system, version 1. *Geoscientific Model Development*, *12*(3), 1241–1266. https://doi.org/10.5194/gmd-12-1241-2019

Pierce, D. W., Das, T., Cayan, D. R., Maurer, E. P., Miller, N. L., Bao, Y., et al. (2013). Probabilistic estimates of future changes in California temperature and precipitation using statistical and dynamical downscaling. *Climate Dynamics*, *40*(3–4), 839–856. https://doi.org/10.1007/s00382-012-1337-9

Randles, C. A., da Silva, A. M., Buchard, V., Colarco, P. R., Darmenov, A., Govindaraju, R., et al. (2017). The MERRA-2 aerosol reanalysis, 1980 onward. Part I: System description and data assimilation evaluation. *Journal of Climate*, *30*(17), 6823–6850. https://doi.org/10.1175/JCLI-D-16-0609.1

Reid, J. S., Benedetti, A., Colarco, P. R., & Hansen, J. A. (2011). International operational aerosol observability workshop. *Bulletin of the American Meteorological Society*, *92*(6), ES21–ES24. https://doi.org/10.1175/2010BAMS3183.1

Reid, J. S., Hyer, E. J., Prins, E. M., Westphal, D. L., Zhang, J., Wang, J., et al. (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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *2*(3), 144–162. https://doi.org/10.1109/JSTARS.2009.2027443

Rolph, G. D., Draxler, R. R., Stein, A. F., Taylor, A., Ruminski, M. G., Kondragunta, S., et al. (2009). Description and verification of the NOAA smoke forecasting system: The 2007 fire season. *Weather and Forecasting*, *24*(2), 361–378. https://doi.org/10.1175/2008WAF2222165.1

Ryan, K. C., Knapp, E. E., & Varner, J. M. (2013). Prescribed fire in North American forests and woodlands: History, current practice, and challenges. *Frontiers in Ecology and the Environment*, *11*(s1), e15–e24. https://doi.org/10.1890/120329

Saide, P. E., Peterson, D. A., da Silva, A., Anderson, B., Ziemba, L. D., Diskin, G., et al. (2015). Revealing important nocturnal and day-to-day variations in fire smoke emissions through a multiplatform inversion. *Geophysical Research Letters*, *42*(9), 3609–3618. https://doi.org/10.1002/2015GL063737

Salguero, J., Li, J., Farahmand, A., & Reager, J. T. (2020). Wildfire trend analysis over the Contiguous United States using remote sensing observations. *Remote Sensing*, *12*(16), 2565. https://doi.org/10.3390/rs12162565

Sampson, C. R., Franklin, J. L., Knaff, J. A., & DeMaria, M. (2008). Experiments with a simple tropical cyclone intensity consensus. *Weather and Forecasting*, *23*(2), 304–312. https://doi.org/10.1175/2007waf2007028.1

Schoennagel, T., Balch, J. K., Brenkert-Smith, H., Dennison, P. E., Harvey, B. J., Krawchuk, M. A., et al. (2017). Adapt to more wildfire in western North American forests as climate changes. *Proceedings of the National Academy of Sciences*, *114*(18), 4582–4590. https://doi.org/10.1073/pnas.1617464114

Schwede, D., Pouliot, G. A., & Pierce, T. (2005). Changes to the Biogenic Emissions Inventory System Version 3 (BEIS3). In *Proceedings of the 4th CMAS models-3 users' conference, Chapel Hill, NC, 26–28 September 2005*.

Sessions, W., Reid, J., Benedetti, A., Colarco, P., Da Silva, A., Lu, S., et al. (2015). Development towards a global operational aerosol consensus: Basic climatological characteristics of the International Cooperative for Aerosol Prediction Multi-Model Ensemble (ICAP-MME). *Atmospheric Chemistry and Physics*, *15*(1), 335–362. https://doi.org/10.5194/acp-15-335-2015

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Liu, Z., Berner, J., & Huang, X.-Y. (2019). A description of the advanced research WRF model Version 4.1 (No. NCAR/TN-556+STR). https://doi.org/10.5065/1dfh-6p97

Sofiev, M., Ermakova, T., & Vankevich, R. (2012). Evaluation of the smoke-injection height from wild-land fires using remote-sensing data. *Atmospheric Chemistry and Physics*, *12*(4), 1995–2006. https://doi.org/10.5194/acp-12-1995-2012

Solazzo, E., Bianconi, R., Vautard, R., Appel, K. W., Moran, M. D., Hogrefe, C., et al. (2012). Model evaluation and ensemble modelling of surface-level ozone in Europe and North America in the context of AQMEII. *Atmospheric Environment*, *53*, 60–74. https://doi.org/10.1016/j.atmosenv.2012.01.003

Spracklen, D. V., Mickley, L. J., Logan, J. A., Hudman, R. C., Yevich, R., Flannigan, M. D., & Westerling, A. L. (2009). Impacts of climate change from 2000 to 2050 on wildfire activity and carbonaceous aerosol concentrations in the western United States. *Journal of Geophysical Research*, *114*(D20), D20301. https://doi.org/10.1029/2008JD010966

Stein, A. F., Draxler, R. R., Rolph, G. D., Stunder, B. J. B., Cohen, M. D., & Ngan, F. (2015). NOAA's HYSPLIT atmospheric transport and dispersion modeling system. *Bulletin of the American Meteorological Society*, *96*(12), 2059–2077. https://doi.org/10.1175/BAMS-D-14-00110.1

Stein, A. F., Rolph, G. D., Draxler, R. R., Stunder, B., & Ruminski, M. (2009). Verification of the NOAA smoke forecasting system: Model sensitivity to the injection height. *Weather and Forecasting*, *24*(2), 379–394. https://doi.org/10.1175/2008WAF2222166.1

Stevens-Rumann, C. S., Kemp, K. B., Higuera, P. E., Harvey, B. J., Rother, M. T., Donato, D. C., et al. (2018). Evidence for declining forest resilience to wildfires under climate change. *Ecology Letters*, *21*(2), 243–252. https://doi.org/10.1111/ele.12889

Tang, Y., Chai, T., Pan, L., Lee, P., Tong, D., Kim, H.-C., & Chen, W. (2015). Using optimal interpolation to assimilate surface measurements and satellite AOD for ozone and $PM_{2.5}$: A case study for July 2011. *Journal of the Air & Waste Management Association*, *65*(10), 1206–1216. https://doi.org/10.1080/10962247.2015.1062439

Theurich, G., DeLuca, C., Campbell, T., Liu, F., Saint, K., Vertenstein, M., et al. (2016). The earth system prediction suite: Toward a coordinated U.S. modeling capability. *Bulletin of the American Meteorological Society*, *97*(7), 1229–1247. https://doi.org/10.1175/BAMS-D-14-00164.1

United States Environmental Protection Agency. (2020a). CMAQ (Version 5.3.2) [Software]. https://doi.org/10.5281/zenodo.4081737

United States Environmental Protection Agency. (2020b). Review of the national ambient air quality standards for particulate matter (pp. 82684–82748). Retrieved From https://www.govinfo.gov/content/pkg/FR-2020-12-18/pdf/2020-27125.pdf

Uprety, S., Cao, C., Xiong, X., Blonski, S., Wu, A., & Shao, X. (2013). Radiometric intercomparison between Suomi-NPP VIIRS and aqua MODIS reflective solar bands using simultaneous nadir overpass in the low latitudes. *Journal of Atmospheric and Oceanic Technology*, *30*(12), 2720–2736. https://doi.org/10.1175/JTECH-D-13-00071.1

Varga, K., Jones, C., Trugman, A., Carvalho, L. M. V., McLoughlin, N., Seto, D., et al. (2022). Megafires in a Warming World: What wildfire risk factors led to California's largest recorded wildfire. *Fire*, *5*(1), 16. https://doi.org/10.3390/fire5010016

Vernon, C. J., Bolt, R., Canty, T., & Kahn, R. A. (2018). The impact of MISR-derived injection height initialization on wildfire and volcanic plume dispersion in the HYSPLIT model. *Atmospheric Measurement Techniques*, *11*(11), 6289–6307. https://doi.org/10.5194/amt-11-6289-2018

Vukovich, J. M., & Pierce, T. (2002). The implementation of BEIS3 within the SMOKE modeling framework.

Xian, P., Reid, J. S., Hyer, E. J., Sampson, C. R., Rubin, J. I., Ades, M., et al. (2019). Current state of the global operational aerosol multi-model ensemble: An update from the International Cooperative for Aerosol Prediction (ICAP). *Quarterly Journal of the Royal Meteorological Society*, *145*(S1), 176–209. https://doi.org/10.1002/qj.3497

Ye, X., Arab, P., Ahmadov, R., James, E., Grell, G. A., Pierce, B., et al. (2021). Evaluation and intercomparison of wildfire smoke forecasts from multiple modeling systems for the 2019 Williams Flats fire. *Atmospheric Chemistry and Physics*, *21*(18), 14427–14469. https://doi.org/10.5194/acp-21-14427-2021

Zhang, H., Kondragunta, S., Laszlo, I., Liu, H., Remer, L. A., Huang, J., et al. (2016). An enhanced VIIRS Aerosol Optical Thickness (AOT) retrieval algorithm over land using a global surface reflectance ratio database. *Journal of Geophysical Research: Atmospheres*, *121*(18), 10717–10738. https://doi.org/10.1002/2016JD024859

Zhang, J., Reid, J. S., Westphal, D. L., Baker, N. L., & Hyer, E. J. (2008). A system for operational aerosol optical depth data assimilation over global oceans. *Journal of Geophysical Research*, *113*(D10), D10208. https://doi.org/10.1029/2007JD009065

Zhang, L., Montuoro, R., McKeen, S. A., Baker, B., Bhattacharjee, P. S., Grell, G. A., et al. (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). *Geoscientific Model Development*, *15*(13), 5337–5369. https://doi.org/10.5194/gmd-15-5337-2022

Zhang, X., Kondragunta, S., Da Silva, A., Lu, S., Ding, H., Li, F., & Zhu, Y. (2019). The blended Global Biomass Burning Emissions Product from MODIS and VIIRS observations (GBBEPx) Version 3.1. Retrieved from https://www.ospo.noaa.gov/Products/land/gbbepx/docs/GBBEPx_ATBD.pdf

Zhang, X., Kondragunta, S., Ram, J., Schmidt, C., & Huang, H.-C. (2012). Near-real-time global biomass burning emissions product from geostationary satellite constellation. *Journal of Geophysical Research*, *117*(D14), D14201. https://doi.org/10.1029/2012JD017459

Zhang, X., Kondragunta, S., & Roy, D. P. (2014). Interannual variation in biomass burning and fire seasonality derived from geostationary satellite data across the contiguous United States from 1995 to 2011. *Journal of Geophysical Research: Biogeosciences*, *119*(6), 1147–1162. https://doi.org/10.1002/2013JG002518

Zhou, X., Zhu, Y., Hou, D., Fu, B., Li, W., Guan, H., et al. (2022). The development of the NCEP global ensemble forecast system Version 12. *Weather and Forecasting*, *37*(6), 1069–1084. https://doi.org/10.1175/waf-d-21-0112.1

Zhu, L., Val Martin, M., Gatti, L., Kahn, R., Hecobian, A., & Fischer, E. (2018). Development and implementation of a new Biomass Burning Emissions Injection Height Scheme (BBEIH v1.0) for the GEOS-Chem model (v9-01-01). *Geoscientific Model Development*, *11*(10), 4103–4116. https://doi.org/10.5194/gmd-11-4103-2018