DEVELOPMENT AND EVALUATION OF NORTH AMERICA ENSEMBLE FORECASTS OF WILDFIRES AND DUST STORMS

by

Peewara Makkaroon
A Thesis
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Master of Science
Earth Systems Science

Committee:

_________________________________________  Dr. Daniel Tong, Thesis Chair

_________________________________________  Dr. Yunyao Li, Committee Member

_________________________________________  Dr. Zafer Boybeyi, Committee Member

_________________________________________  Dr. Mark Uhen, Department Chairperson

_________________________________________  Dr. Donna M. Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science

_________________________________________  Dr. Fernando R. Miralles-Wilhelm
Dean, College of Science

Date:  ______________________________ Spring Semester 2022
George Mason University
Fairfax, VA
Development and Evaluation of North America Ensemble Forecasts of Wildfires and Dust Storms

A Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at George Mason University

by

Peewara Makkaroon
Bachelor of Science
Chulalongkorn University, 2018

Director: Daniel Tong, Associate Professor
Department of Atmospheric, Oceanic, and Earth Sciences

Spring Semester 2022
George Mason University
Fairfax, VA
DEDICATION

This thesis is dedicated to my loving family and friends, who have always encouraged me and supported me throughout the challenging degree program.
ACKNOWLEDGEMENTS

I would like to express my most profound appreciation to my committee chair, Dr. Daniel Tong. He gave me an opportunity to work as a graduate research assistant at George Mason University. He constantly shares his research and work experiences with me and guides me to complete this thesis successfully. I would also like to thank my two committee members, Dr. Yunyao Li, who always shares research and coding resources with me, and Dr. Zafer Boybeyi, who provides valuable recommendations on my thesis. I will always appreciate all they have done. Finally, I’m thankful to my academic advisor, Dr. Geoffrey Gilleaudeau, for guiding me with class registration and thesis processes during the degree program.

This study is financially supported by NASA Health and Air Quality Program and NOAA Weather Program Office. We thank NASA, NOAA, and NRL for providing the model prediction data to create the ensemble forecast. In addition, ground measurements collected by the EPA and satellite AOD data by NOAA and NASA are gratefully acknowledged.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Tables</td>
<td>vii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>viii</td>
</tr>
<tr>
<td>Abstract</td>
<td>x</td>
</tr>
<tr>
<td>Chapter 1: Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Wildfires in the United States</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Dust Storms in the United States</td>
<td>2</td>
</tr>
<tr>
<td>1.3 Multi-Model Ensemble Forecasts</td>
<td>4</td>
</tr>
<tr>
<td>1.4 Objectives</td>
<td>5</td>
</tr>
<tr>
<td>Chapter 2: Case Studies of Wildfires and Dust Storms</td>
<td>8</td>
</tr>
<tr>
<td>2.1 The 2020 Gigafire in the Western United States</td>
<td>8</td>
</tr>
<tr>
<td>2.2 The 2021 Spring Dust Season in the Chihuahuan Desert</td>
<td>10</td>
</tr>
<tr>
<td>CHAPTER 3: Description of Observations</td>
<td>12</td>
</tr>
<tr>
<td>3.1 Ground-Based Observation</td>
<td>12</td>
</tr>
<tr>
<td>3.1.1 AirNow PM$_{2.5}$</td>
<td>12</td>
</tr>
<tr>
<td>3.2 Satellite-Based Observations</td>
<td>12</td>
</tr>
<tr>
<td>3.2.1 MAIAC AOD</td>
<td>12</td>
</tr>
<tr>
<td>3.2.2 VIIRS-SNPP AOD</td>
<td>13</td>
</tr>
<tr>
<td>Chapter 4: Multi-Model Ensemble Creation</td>
<td>15</td>
</tr>
<tr>
<td>4.1 Description of Ensemble Members</td>
<td>15</td>
</tr>
<tr>
<td>4.1.1 GMU-CMAQ</td>
<td>15</td>
</tr>
<tr>
<td>4.1.2 NACC-CMAQ</td>
<td>16</td>
</tr>
<tr>
<td>4.1.3 HYSPLIT</td>
<td>17</td>
</tr>
<tr>
<td>4.1.4 GEFS-Aerosols</td>
<td>17</td>
</tr>
<tr>
<td>4.1.5 GEOS-5</td>
<td>18</td>
</tr>
<tr>
<td>4.1.6 ICAP-MME</td>
<td>19</td>
</tr>
<tr>
<td>4.1.7 NAAPS</td>
<td>20</td>
</tr>
</tbody>
</table>

v
4.2 Multi-Model Ensemble Forecasts ............................................................... 21
  4.2.1 Ensemble AOD Forecasts ................................................................. 21
  4.2.2 Ensemble PM\textsubscript{2.5} Forecasts ........................................... 22
  4.3 Ensemble Probability of PM\textsubscript{2.5} Exceedance Forecast ............. 22

Chapter 5: Evaluation Methodology ................................................................. 24
  5.1 Statistical Metrics ............................................................................... 24
  5.2 Categorical Metrics ........................................................................... 25
  5.3 Overall Rating ..................................................................................... 27

Chapter 6: Evaluation of Ensemble Forecast of Wildfires ................................. 29
  6.1 Ensemble Performance in Forecasting AOD ....................................... 29
  6.2 Ensemble Performance in Forecasting Surface PM\textsubscript{2.5} Concentration .... 31
  6.3 Overall Performance of Ensemble Wildfire Forecast ......................... 33
  6.4 Ensemble Probability Forecast of PM\textsubscript{2.5} Exceedances ................. 42

Chapter 7: Evaluation of Ensemble Forecast of Dust Storms ............................. 47
  7.1 Ensemble AOD Forecasting Performance on March 16\textsuperscript{th}, 2021 .... 47
  7.2 Ensemble PM\textsubscript{2.5} Forecasting Performance on March 16\textsuperscript{th}, 2021 54
  7.3 Ensemble Forecasting Performance during the 2021 Spring Dust Season .... 58
  7.4 Overall Performance of Ensemble Dust Storm Forecasts ..................... 60
  7.5 Ensemble Probability Forecast of PM\textsubscript{2.5} Exceedances .................... 67

Chapter 8: Conclusions ................................................................................... 73
  8.1 Conclusion and Recommendation for Future Work ............................ 73

Appendix: The configuration of participating model ....................................... 78

References ..................................................................................................... 79
Biography ....................................................................................................... 99
LIST OF TABLES

Table 1. Overall ensemble mean and individual model performances in forecasting AOD values and PM$_{2.5}$ concentrations during the 2020 Gigafire events..................42
Table 2. Averaged $aH$ and $aFAR$ values of ensemble probability of PM$_{2.5}$ exceedance forecast during the 2020 Gigafire events.........................................................46
Table 3. Correlation between AOD simulations and VIIRS Deep Blue (DB) AOD observations over the Contiguous United States (CONUS) and Active Dust Regions in western Texas and southern New Mexico on March 16$^{th}$, 2021..........................53
Table 4. Overall ensemble mean and individual model performances in forecasting AOD values and PM$_{2.5}$ concentrations on March 16$^{th}$, 2021........................................54
Table 5. Overall ensemble mean and individual model performances in forecasting AOD values and PM$_{2.5}$ concentrations during the 2021 Spring Dust Season............67
Table 6. Averaged $aH$ and $aFAR$ values of ensemble probability of PM$_{2.5}$ exceedance forecast during the 2021 Spring Dust Season..............................................72
Table A1. The configuration of participating models included in the ensemble forecasting........................................................................................................78
<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1. VIIRS-SNPP true color imagery overlaid by PM$<em>{2.5}$ observations measured by AQS sites on September 12$^{th}$, 2020. The time series plot of daily maximum PM$</em>{2.5}$ concentrations measured by all AirNow sites across the CONUS during the Gigafire events.</td>
<td>9</td>
</tr>
<tr>
<td>Figure 2. VIIRS/SNPP true color imagery overlaid by PM$<em>{2.5}$ observations measured by AQS sites in western Texas and southern New Mexico on March 16$^{th}$, 2021. The time series plot of daily maximum PM$</em>{2.5}$ concentrations measured by AirNow sites in the southwestern United States during the 2021 Spring Dust Season.</td>
<td>11</td>
</tr>
<tr>
<td>Figure 3. Example of scatter plot for definition of categorical metrics</td>
<td>27</td>
</tr>
<tr>
<td>Figure 4. AOD predicted by seven individual models (a-g) and the ensemble mean (h), compared with VIIRS enhanced Dark Target (DT) AOD retrievals (i) on August 22, 2020.</td>
<td>35</td>
</tr>
<tr>
<td>Figure 5. AOD predicted by seven individual models (a-g) and the ensemble mean (h), compared with MAIAC AOD retrievals (i) on August 22, 2020.</td>
<td>36</td>
</tr>
<tr>
<td>Figure 6. Surface PM$<em>{2.5}$ concentrations predicted by six individual models (a-f) and the ensemble mean (g), compared with AirNow PM$</em>{2.5}$ observations (h) on August 22, 2020.</td>
<td>37</td>
</tr>
<tr>
<td>Figure 7. Time series of RMSE (a), CORR (b), MB (c), ME (d), NMB (e), NME (f), and FB (g) of AOD for the 2020 Gigafire events.</td>
<td>38</td>
</tr>
<tr>
<td>Figure 8. Time series of RMSE (a), CORR (b), MB (c), ME (d), NMB (e), NME (f), and FB (g) of AOD for the 2020 Gigafire events.</td>
<td>39</td>
</tr>
<tr>
<td>Figure 9. Time series of RMSE (a), CORR (b), MB (c), ME (d), NMB (e), NME (f), aH (g), aFAR (h), and FB (i) of PM$_{2.5}$ for the 2020 Gigafire events.</td>
<td>40</td>
</tr>
<tr>
<td>Figure 10. Time series of the overall rating (RANK) for AOD and PM$_{2.5}$ simulated by the ensemble mean and individual models during the 2020 Gigafire events.</td>
<td>41</td>
</tr>
<tr>
<td>Figure 11. Ensemble probability forecast of PM$_{2.5}$ exceedances on August 22$^{nd}$, 2020 (during the 2020 Gigafire events).</td>
<td>45</td>
</tr>
<tr>
<td>Figure 12. Time series plots of aH (a) and aFAR (b) values during the 2020 Gigafire events for the ensemble probability of PM$_{2.5}$ exceedance forecast.</td>
<td>46</td>
</tr>
<tr>
<td>Figure 13. AOD predicted by six individual models (a-f) and the ensemble mean (g), compared with VIIRS enhanced Dark Target (DT) AOD retrievals (h) and VIIRS Deep Blue (DB) AOD retrievals (i) on March 16$^{th}$, 2021.</td>
<td>51</td>
</tr>
<tr>
<td>Figure 14. AOD predicted by six individual models (a-f) and the ensemble mean (g), compared with VIIRS Deep Blue (DB) AOD retrievals (h) near the active dust regions in western Texas and southern New Mexico on March 16$^{th}$, 2021.</td>
<td>52</td>
</tr>
</tbody>
</table>
Figure 15. Surface PM$_{2.5}$ concentrations predicted by five individual models (a-e) and the ensemble mean (f), compared with AirNow PM$_{2.5}$ observations (g) on March 16$^{th}$, 2021.

Figure 16. Time series of RMSE (a), CORR (b), MB (c), ME (d), NMB (e), NME (f), and FB (g) of AOD during the 2021 Spring Dust Season.

Figure 17. Time series of RMSE (a), CORR (b), MB (c), ME (d), NMB (e), NME (f), and FB (g) of AOD during the 2021 Spring Dust Season.

Figure 18. Time series of RMSE (a), CORR (b), MB (c), ME (d), NMB (e), NME (f), aH (g), aFAR (h), and FB (i) of PM$_{2.5}$ for the 2021 Spring Dust Season during January-March 2021.

Figure 19. Time series of the overall rating (RANK) for AOD and PM$_{2.5}$ simulated by the ensemble mean and individual models during the 2021 Spring Dust Season.

Figure 20. Ensemble probability forecast of PM$_{2.5}$ exceedances on March 16$^{th}$, 2021 (during the 2021 Spring Dust Season).

Figure 21. Time series plots of aH (a) and aFAR (b) values during the 2021 Spring Dust Season (January-March 2021) for the ensemble probability of PM$_{2.5}$ exceedance forecast.
Wildfires and dust storms are two major emission sources of aerosols in the atmosphere, exerting myriad effects on air quality, climate, and human health. Predicting wildfires and dust storms is challenging due to large uncertainties in the inputs and representation of chemical and physical processes in the atmospheric models. Ensemble forecasting has been proposed to improve the predictability of wildfire and dust aerosols. This work presents the development and evaluation of a multi-model ensemble forecast system of wildfire and dust air pollution over North America, leveraging research and operational forecasts operated by George Mason University (GMU) and three U.S. federal agencies: National Oceanic and Atmospheric Administration (NOAA), National Aerospace and Space Agency (NASA), and Naval Research Laboratory (NRL). The ensemble members include three regional models (GMU CMAQ, NOAA NACC-CMAQ, and NOAA HYSPLIT), three global models (NOAA GEFS-Aerosols, NASA GEOS-5, and NRL
Performance of the ensemble forecast was evaluated with aerosol optical depth (AOD) products from MODIS MAIAC, VIIRS-SNPP enhanced Dark Target (DT) and Deep Blue (DB), and surface PM$_{2.5}$ (fine particle) from the AirNow ground network during the 2020 Gigafire events (August-September 2020) in the western United States and the 2021 Spring Dust Season in the Chihuahuan Desert.

For the wildfire ensemble, the results showed that, compared to the individual models, the ensemble mean significantly reduced the biases in the wildfire air pollution forecasts and produced more persistently reliable forecasts during extreme fire events. For AOD forecasts, the ensemble mean was able to improve model performance, such as increasing the correlation to 0.57 (0.62) from a range of 0.30-0.53 (0.35-0.56) by individual models when compared to the VIIRS (MAIAC). The ensemble mean also yields the best (second best) overall $RANK$, a composite indicator representing four statistical metrics (correlation, fractional bias, area hit rate, and false alarm ratio) compared to VIIRS (MAIAC). For the forecast of surface PM$_{2.5}$ concentration, the ensemble mean demonstrated better performance than any single model 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), when compared to the AirNow observations. Finally, the ensemble shows the potential to provide a suitable exceedance probability forecast during wildfires with the lowest area false alarm ratio (1.52%) achieved by the ensemble probability of 100%.
For the dust ensemble, the ensemble mean moderately reduced biases in the dust air pollution forecasts and provided fairly reliable AOD and PM$\textsubscript{2.5}$ forecasts during extreme dust storms compared to the individual models. For AOD forecasts, the ensemble mean improved forecasting performance less successfully than expected, as demonstrated by slightly decreasing mean bias to 0.01 (0.07) based on VIIRS DT (VIIRS DB), increasing correlation to 0.32 at the low level highest from a range of 0.09-0.31 (VIIRS DB), and yielding the third best overall $RANK$ compared to VIIRS DT and DB. For surface PM$\textsubscript{2.5}$ forecasts, the ensemble mean underperformed with a slightly reduced mean bias (3.14), moderately improved low-level correlation (0.40), low area hit rates (15%), and the third best $RANK$. The ensemble was able to provide only low-medium (20-60%) exceedance probability forecasts during dust events. In addition, the low correlations and large biases of the dust ensemble forecasts during the extreme dust episodes indicate worse performance compared to that of wildfire ensemble forecasts due to larger uncertainties in predicting dust emission, dispersion, and removal.

The thesis findings highlight that using the ensemble approach can reduce biases in air pollution forecasts and reasonably improve the model predictability during extreme events such as wildfires and dust storms. The proposed ensemble exceedance probability forecast can be further applied to early warnings of severe air pollution episodes during wildfires and dust storms. However, the reliability of the ensemble forecast is still subject to types of extreme events due to different emission sources as well as initial and boundary meteorological conditions.
CHAPTER 1: INTRODUCTION

1.1 Wildfires in the United States

Wildfires are important emission sources that contribute large amounts of aerosols and trace gases to the atmosphere, leading to hazardous air quality. Wildfire air pollution causes adverse respiratory health effects, visibility degradation, and premature mortality, which in turn lead to economic burdens (Fann et al., 2018; Ford et al., 2018; Neumann et al., 2021). 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. (Li and Banerjee, 2021; McClure and Jaffe, 2018; 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., more than 10 million acres burned (NICC, 2020), and that most fires took place in the Western United States. Northern California in particular was affected and has experienced the
largest recorded wildfires during Summer-Fall 2020 fire season (California Department of Forestry and Fire Protection [CAL FIRE], 2020).

1.2 Dust Storms in the United States

Dust particles are known as a major component of particulate matter less than 2.5 \( \mu \text{m} \) in aerodynamic diameter (PM\(_{2.5}\)) in the western U.S. during spring (Hand et al., 2011, 2016, 2017) as a result of powerful, sustained winds and relatively low precipitation (Flagg et al., 2014). Dust events during spring and summer have been on the rise in the western U.S. over the past several decades (Tong et al., 2017), and it will be worsening significantly in the upcoming years over the southern Great Plains due to severe drought, soil moisture deficits, warming temperatures, and variations in sea surface temperature from climate change (Achakulwisut et al., 2018; Hand et al., 2016; Pu and Ginoux, 2017, 2018; Tong et al., 2017). This projected trend has drawn many concerns about its detrimental impacts on the atmospheric environment (Balkanski et al., 2007; Benedetti et al., 2014; Forster et al., 2007; Wu et al., 2016), ecosystem (Barkley et al., 2019; Mills, 2004; Prospero et al., 2020; Swap et al., 1994), and human health (World Health Organization [WHO], 2021). Dust storms can lift large amounts of soil-derived dust particles into the air. As a result, the concentrations of small particulate matter within active dust regions are elevated beyond the safety air quality standard level and can induce adverse health effects, such as severe respiratory diseases (Tobias et al., 2019), cardiovascular health issues (Crooks et al., 2016), as well as raising Valley Fever incidence rate (CDC, 2013; Tong et al., 2017). Apart from the direct health impacts, dust
particles contribute to increasing transportation accidents due to degraded visibility, especially on highways during intense dust storms (Ashley et al., 2015; Lader et al., 2016; Van Pelt et al., 2020b).

Dust in the western U.S. is primarily generated by wind erosion of exposed soil surfaces in arid or semi-arid regions and can be transported across the Contiguous United States (CONUS). Natural sources of dust in the western U.S. are the North American Deserts (namely the Chihuahuan, Great Basin, Mojave, and Sonoran deserts) (Ginoux et al., 2012; Jewell & Nicoll, 2011; Reynolds et al., 2007; Rivera et al., 2010; Tanaka & Chiba, 2006). In contrast, anthropogenic sources of dust are primarily associated with agricultural activities in the southern Great Plains and the Colorado Plateau (Carmona et al., 2015; Ginoux et al., 2012; Neff et al., 2008; Saxton et al., 2000; Skiles et al., 2015; Reynolds et al., 2016). In addition to local dust sources in the U.S., long-range dust transports from Asia and Africa across the Atlantic and Pacific Oceans in spring and summer (March-August) subsequently contribute to total dust in the United States. Generally, the trans-Pacific dust, transported from Asia by strong tropical cyclones and westerly winds during spring (February to June), frequently affects the Pacific coastal regions of the western U.S. (Creamean et al., 2014; Fairlie et al., 2007; Fischer et al., 2009; Kavouras et al., 2009; VanCuren & Cahill, 2002; Zhao et al., 2008), while the trans-Atlantic dust transported from Africa by powerful easterly winds commonly impacts the Caribbean Islands, the Gulf of Mexico, and the southeastern U.S. (Prospero, 1981; Prospero and MayolBracero, 2013; Prospero and colleagues, 2021). These long-range transported dusts contribute to high background particulate matter (PM)
concentrations not in both the western U.S. (Fischer et al., 2009; Jaffe et al., 2003) and eastern (DeBell et al., 2004) during spring and summer.

1.3 Multi-Model Ensemble Forecasts

Regarding the concerns about human health affected by degraded air quality during wildfires and dust storms, many operational forecasting systems have been developed to forecast the dispersion of aerosols with the main goal of protecting the public from harmful air quality during hazardous air quality events (Basart et al., 2012; Campbell et al., 2021; Colarco et al., 2010; Hamill et al., 2011a, b; Johnson et al., 2011; Liu et al. 2007; Lu et al. 2010, 2013; Liu and Westphal 2001; Marticorena and Bergametti, 1995; Marticorena et al., 1997; Nickovic et al. 2001; Pérez et al. 2011; Rienecker et al. 2008; Stein et al., 2015; Terradellas et al. 2011; Wang et al., 2000; Walker et al., 2009; Xian et al., 2019; Li et al., 2021). However, the accuracy of deterministic forecasts from a single model is predominantly deteriorated by uncertainties in emission and meteorological input data, model simulations, physical and chemical processes (Cakmur et al., 2004; Darmenova et al., 2009; Delle Monache and Stull, 2003; Di Tomaso et al., 2017; Ginoux et al., 2012; Gong and Zhang, 2008; Grini et al., 2005; Kang et al., 2011; Kumar et al., 2020; Marticorena and Bergametti, 1995; Li et al., 2020; Shao et al., 1996; Textor et al., 2006; Uno et al., 2006), and surface properties (e.g., soil roughness, soil moisture, and vegetation types) (Grini et al., 2005).

Alternatively, one effective way to improve predicting performance is using a mean of the ensemble approach, which can provide probabilistic forecasts by calculating
the mean from either multiple models or input data (Delle Monache and Stull, 2003; Delle Monache et al. 2006a, b, 2008; Delle Monache et al., 2020; Li et al., 2020; Petersen et al., 2019; Solazzo et al., 2012; Xian et al., 2019). The major advantage of the ensemble mean forecast over a single model forecast is that it can reduce the biases in forecasts of ensemble members by averaging them out and the uncertainties in ensemble forecasts can also be determined from the spreads of ensemble members.

1.4 Objectives

This thesis aims to develop multi-model ensemble forecasts of wildfire and dust air pollution based on the mean of participating models (ensemble mean) for the Contiguous United States (CONUS). The ensemble members include three regional models, three global models, and one global ensemble. The regional systems include the George Mason University-Community Multiscale Air Quality (GMU-CMAQ), National Oceanic and Atmospheric Administration-U.S. Environmental Protection Agency (NOAA-EPA) Atmosphere-Chemistry Coupler-Community Multiscale Air Quality (NACC-CMAQ), and NOAA Hybrid Single-Particle Lagrangian Integrated Trajectory (HYSPLIT) models. GMU-CMAQ is a research forecasting system run by the air quality group of George Mason University (GMU) (Li et al., 2021) to provide daily air quality forecasts across the U.S. for the general public. NACC-CMAQ, a model currently being used in NOAA's operational National Air Quality Forecasting Capability (NAQFC) (Campbell et al., 2022), and HYSPLIT, a common atmospheric transport and dispersion model that is developed at NOAA/Air Resources Laboratory (ARL) and used widely in
the atmospheric sciences community (Stein et al., 2015). The four global models are: Global Ensemble Forecast System Aerosols (GEFS-Aerosols), NASA Goddard Earth Observing System (GEOS, version 5), International Cooperative for Aerosol Prediction Multi-Model aerosol forecasting Ensemble (ICAP-MME), and Navy Aerosol Analysis and Prediction System (NAAPS). GEFS-Aerosols, a global atmospheric composition model, is developed by at the National Centers for Environmental Prediction (NCEP) in collaboration with the NOAA Global Systems Laboratory (GSL), NOAA Chemical Sciences Laboratory (CSL), and NOAA/ARL (Hamill et al., 2011a, b). GEOS is a weather and climate capable model and is a significant part of the GEOS atmospheric data assimilation system (DAS) and Earth system model developed at NASA’s Global Modeling and Assimilation Office (GMAO) (Rienecker et al., 2008). NAAPS, the U.S. Navy’s operational global aerosol transport model with consideration of processes associated with aerosol lifecycles, and AOD data assimilation, is developed at the Naval Research Laboratory (NRL) (Lynch et al., 2016). Finally, the ICAP-MME is a global ensemble mean produced from nine comprehensive global speciated aerosol and/or dust models (Xian et al., 2019).

The performance of the ensemble mean in forecasting Aerosol Optical Depth (AOD) were intercompared with ensemble members and verified by evaluating AOD simulations against the Multi-Angle Implementation of Atmospheric Correction (MAIAC; Lyapustin et al., 2011a,b; 2012; 2018) and Visible Infrared Imaging Radiometer Suite onboard the Suomi National Polar-orbiting Partnership (SNPP) (VIIRS-SNPP; Cao et al., 2013a, 2013b; Uprety et al., 2013) satellite-retrieved AOD.
products based on enhanced Dark Target (DT) algorithm over dark and bright surfaces (Zhang et al., 2016) for the application to the wildfires, while the AOD simulations were evaluated against the VIIRS enhanced DT and VIIRS Deep Blue (DB; Hsu et al., 2013; Hsu et al., 2019) products for the application to the dust storms. The PM$_{2.5}$ concentrations simulated by the ensemble mean were also intercompared with ensemble members and verified with the AirNow PM$_{2.5}$ ground observations for both applications to wildfires and dust storms. The evaluation results were derived by analyzing a suite of statistical metrics during the 2020 Gigafire event (August-September 2020), caused by the August Complex Fire burning more than 1 million acres in Northern California, and during the 2021 Spring Dust Season (January-March 2021), a period when the dust storms driven by a strong low-pressure system occurred predominantly in the Chihuahuan Desert in western Texas and southern New Mexico.

As the air quality models are used to provide air pollution warnings to the public, the ability of the ensemble to produce a reliable forecast of health-based PM$_{2.5}$ exceedances of NAAQS (24-hr PM$_{2.5}$ concentration above 35 µg/m$^3$; U.S. EPA, 2020a) during extreme wildfires is crucial. Therefore, we created the ensemble probability forecasts of PM$_{2.5}$ exceedances influenced by wildfires and dust storms and evaluated them with the observed exceedances by AirNow ground monitoring network.
CHAPTER 2: CASE STUDIES OF WILDFIRES AND DUST STORMS

2.1 The 2020 Gigafire in the Western United States

In 2020, California experienced 9,917 incidents of multiple complex wildfires, leading to over 4 million acres burned, and 10,488 structures destroyed (CAL FIRE, 2020). On August 16th, the largest and the most complex recorded wildfire ever known as the “August Complex fire” occurred in Northern California. The fire initially started burning in the Mendocino National Forest from lightning strikes coupled with a heatwave and severe drought driven by climate change. The August Complex fire was a combination of the Doe, Tatham, Glade, and Hull fires. On September 9th, the Doe fire (main fire) became the single largest complex wildfire, even larger than the 2018 Mendocino Complex fire. Later on, the Doe fire merged with other following fires and continued burning until November 12th. The fire was the first “Gigafire”, active for 86 days, burned more than 1 million acres, and destroyed 935 structures across the Coast Range counties (Colusa, Glenn, Humboldt, Lake, Mendocino, Tehama, and Trinity). The spanning of fires is dominated by the Diablo winds (offshore winds) over these areas. Figure 1a displays extremely high observed PM$_{2.5}$ concentrations from AQI sites mainly in the western U.S. on September 12th, 2021, when the fires were very intense. Extremely high daily PM$_{2.5}$ concentrations above the daily National Ambient Air Quality Standards (NAAQS) for PM$_{2.5}$ (>35 μg/m$^3$) were recorded at many AirNow monitoring sites across
the U.S. between September 10\textsuperscript{th}-17\textsuperscript{th}, 2020 primarily over California, Oregon, and Washington as shown in Figure 1b. Consequently, our study will focus on AOD and PM\textsubscript{2.5} simulations during the 2021 wildfire season, from August to September 2020.

Figure 1. VIIRS-SNPP true color imagery overlaid by PM\textsubscript{2.5} observations measured by AQS sites (circles) on September 12\textsuperscript{th}, 2020, from NOAA AerosolWatch\textsuperscript{1} (above). The time series plot of daily maximum PM\textsubscript{2.5} concentrations measured by all AirNow sites across the Contiguous United States during the Gigafire events from August to September 2020 (bottom).

\textsuperscript{1}https://www.star.nesdis.noaa.gov/smcd/spb/aq/AerosolWatch/
2.2 The 2021 Spring Dust Season in the Chihuahuan Desert

The spring dust season in the western and southwestern United States commonly occurs from January to March. According to the NOAA’s National Weather Service (NWS/NOAA), during Spring 2021, many dust storms occurred primarily over the Chihuahuan Desert from Mexico to across the border in western Texas and southern New Mexico. In the middle of March, from March 13\textsuperscript{th} to 18\textsuperscript{th}, a powerful low-pressure system coupled with drought and the La Niña effect generated gusty winds of 35 to 45 mph across the Mexico-United States border and scattered dust from the Chihuahuan Desert in Mexico to western Texas and southern New Mexico. Consequently, daily average PM\textsubscript{2.5} concentrations over the active dust regions were substantially increased to 50-60 µg/m\textsuperscript{3}, as shown in Figure 2b. On March 16\textsuperscript{th}, a dust storm occurred for nearly eight hours in El Paso, Texas\textsuperscript{2} (Figure 2a), which was the most unusual long-lasting dust storm in the city's history and led to worsened air quality and a decreased visibility of less than a half-mile over El Paso and Juárez, Texas.

\textsuperscript{2} https://earthobservatory.nasa.gov/images/148057/long-lasting-dust-storm-from-chihuahua
Figure 2. VIIRS/SNPP true color imagery overlaid by PM$_{2.5}$ observations measured by AQS sites (circles) in western Texas and southern New Mexico on March 16$^{th}$, 2021, from NOAA AerosolWatch (above). The time series plot of daily maximum PM$_{2.5}$ concentrations measured by AirNow sites in the southwestern United States during the 2021 Spring Dust Season (bottom).
CHAPTER 3: DESCRIPTION OF OBSERVATIONS

3.1 Ground-Based Observation

3.1.1 AirNow PM$_{2.5}$

Hourly PM$_{2.5}$ observations were obtained from the U.S. EPA AirNow network\(^3\). 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 derived from each of AirNow sites, starting from 12:00 UTC of the current day to 11:00 UTC the next day, were averaged into a daily value grid by grid.

3.2 Satellite-Based Observations

3.2.1 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., 2011a, b; 2012; https://www.AirNow.gov

\(^3\)https://www.AirNow.gov
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 daily global 1 km MAIAC AOD at 550 nm from all orbits available for the CONUS, later averaged at each grid location. MAIAC data were provided by NASA GSFC.

### 3.2.2 VIIRS-SNPP AOD

VIIRS-SNPP AOD product was acquired from the VIIRS instrument carried onboard the Suomi National Polar-orbiting Partnership (SNPP), which is a part of the Joint Polar Satellite System (JPSS) (Cao et al., 2013a, 2013b; Uprety et al., 2013). The VIIRS instrument was initially developed based on the previous series of measurements on NOAA satellites and MODIS on the Terra and Aqua satellites (Levy et al., 2013, 2015) through the cooperation of NASA and NOAA. The VIIRS 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 in order to support NASA's Earth Observing System (EOS) and NOAA’s polar-orbiting operational environmental satellite system (POES). 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 (DT) over dark and bright surfaces (Zhang et al., 2016) daily AOD product at 550 nm with a fixed grid resolution of 0.1°× 0.1° as provided by NOAA, and VIIRS-SNPP Level 2 Deep Blue (DB; Hsu et al.,
2013; Hsu et al., 2019; Sayer et al., 2018) 6-minute AOD product at 550 nm with an at-nadir resolution of 6 km x 6 km as provided by NASA GSFC.
CHAPTER 4: MULTI-MODEL ENSEMBLE CREATION

4.1 Description of Ensemble Members

In this section, each of seven participating numerical air quality models included in the ensemble will be described. Model configurations are shown in Table A1 in the Appendix.

4.1.1 GMU-CMAQ

GMU-CMAQ (Li et al., 2021) uses meteorological fields derived from the Weather Research and Forecasting model version 4.2 (WRFv4.2) (Skamarock et al., 2019) to drive the offline CMAQ model version 5.3.1 (CMAQv5.3.1) (US EPA, 2020b), and uses biomass burning (BB) emission data from the Global Biomass Burning Emissions Product (GBBEPx; Zhang et al., 2012, 2014, 2019) blended between Moderate Resolution Imaging Spectroradiometer (MODIS) on the NASA Terra and Aqua satellites and VIIRS-SNPP. 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 (SMOKE) 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, and dust plumes are calculated using the Sofiev et al. (2012) and the FENGCHA dust scheme developed by NOAA/ARL (Dong et al., 2016), respectively.
GMU-CMAQ provides hourly experimental AOD and PM$_{2.5}$ concentration forecasts on a horizontal resolution of 12 km $\times$ 12 km over the CONUS with each day’s forecast initialized at 18:00 UTC on the previous day.

4.1.2 NACC-CMAQ

NACC-CMAQ meteorological preprocessor was adapted from the EPA’s Meteorology Interface Processor (MCIP) version 5 (e.g., NACC version 1.3.2; https://zenodo.org/record/5507489#.YmvzsejMKUk, last access 29 Apr 2022), and uses meteorological fields from NOAA’s latest operational Finite Volume Cubed-Sphere (FV3) Global Forecast System version 16 (GFSv16) to drive the offline CMAQv5.3.1 (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 and Pierce, 2002; Schwede, 2005) with the Biogenic Emission Landuse Dataset version 5 (BELD5) for biogenic volatile organic carbon (BVOC) emissions. The wildfire smoke plumes are computed using the Briggs (1969) plume rise algorithm. The dust plumes are computed using dust algorithms including, the FENGCHA dust scheme, SoilGrids soil fractions (Hengl et al., 2017), surface roughness from merged satellite microwave backscattering (ASCAT), and visible/near-infrared reflectances (PARASOL) (Prigent et al., 2012).

NACC-CMAQ uses meteorology and emission inputs together with aerosol boundary conditions from NOAA’s operational GEFS-Aerosols model for dust and smoke to provide hourly AOD and PM$_{2.5}$ forecasts at a horizontal resolution of 12 km $\times$ 12 km
(same as GMU-CMAQ) with each day’s forecast initialized at 12:00 UTC on of the previous day over CONUS.

4.1.3 HYSPLIT

HYSPLIT (Stein et al., 2015) uses a plume-following coordinate system and back trajectory analysis, and is typically used to determine the emission sources, atmospheric transport, dispersion, deposition, and chemical transformation of aerosols over the CONUS (Draxler & Hess, 1998). 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). HYSPLIT has been recently updated to version 5.1.0 (HYSPLITv5.1.0) and combines WRF-ARW (Advanced Research WRF) meteorology inputs, fire emission products from United States Forest Service (USFS) BlueSky, and Briggs (1969) plume rise scheme to simulate hourly AOD and PM$_{2.5}$ concentration forecasts at a horizontal resolution of $0.15^\circ \times 0.15^\circ$ with each day’s forecast initialized at 00:00 UTC on of the previous day over CONUS.

4.1.4 GEFS-Aerosols

NOAA’s 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 (NUOPC) Layer (Theurich et al., 2016), which is the current and future foundation of NOAA’s Unified Forecast System (UFS) modeling framework (Hamill et al., 2011a, b; L. Zhang et al., 2021). The operational GEFS-Aerosols model
currently uses BB emission data from GBBEPx, and global anthropogenic emission data
from the Community Emission Data System (CEDS) in 2014 for gaseous emissions and
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). The dust plumes are computed using the FENGCHA dust scheme, SoilGrids soil
fractions, surface roughness from merged satellite microwave backscattering (ASCAT),
and visible/near-infrared reflectances (Prigent et al., 2012). 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.

4.1.5 GEOS-5

GEOS is a global data assimilation and forecasting system that combines the
GMAO modified gridpoint statistical interpolation (GSI) analysis algorithm, which was
originally developed by the National Centers for Environment Prediction (NCEP)
Environmental Modeling Center, with the NASA atmospheric global forecast model
(Rienecker et al., 2008). The GEOS version 5.27.1 (GEOSv5.27.1) is integrated using the
Earth System Modeling Framework (ESMF), and its configuration includes
meteorological data acquired by the GEOS Data Assimilation System (DAS) in near real
time, fire detection information from MODIS, emissions of aerosols, BB, and, smoke
data from the Quick Fire Emissions Dataset (QFED), anthropogenic emissions from the
Emissions Database for Global Atmospheric Research (EDGAR)-HTAP inventories, and
Model of Emissions of Gases and Aerosols from Nature (MEGAN) for BVOC emissions.
The dust plumes were computed using the Goddard Global Ozone Chemistry Aerosol Radiation and Transport (GOCART; Colarco et al., 2010) model. This study used GEOS-5 global forecast of hourly AOD values and PM$_{2.5}$ concentrations on a horizontal resolution of $0.25^\circ \times 0.3125^\circ$ and initialized each day at 00:00 UTC.

### 4.1.6 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., 2014a; 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), which provides a testbed of probabilistic aerosol forecasts. ICAP-MME is generated by combining nine global aerosol models: the European Centre 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 (JMA) Model of Aerosol Species in the Global Atmosphere (MASINGAR), NOAA Environmental Modeling System (NEMS) Global Forecast System (GFS) Aerosol Component (NGAC), Météo-France Modèle de Chimie Atmosphérique à Grande Echelle (MOCAGE), and Finnish Meteorological Institute (FMI) System for Integrated-modeLling of Atmospheric coMposition (SILAM), the Barcelona Supercomputing Center (BSC) Chemical Transport Model (CTM), embedded in the Multiscale Online Nonhydrostatic AtmospheRe CHemistry (MONARCH) and the UK Met Office (UKMO) models. These models have different underlying meteorological
fields, emissions, microphysics, and chemistry, as well as a variety of horizontal and vertical resolutions ranging from $0.25^\circ \times 0.31^\circ$ and 72 vertical layers to $1.4^\circ \times 1^\circ$ and 24 layers. As a result, ICAP-MME is driven by the independent operation/quasi-operational meteorology inputs and aerosol variables generated by each of the member organizations. This study utilized ICAP-MME global 6-hour AOD at 550 nm on a horizontal resolution of $1^\circ \times 1^\circ$ and initialized each day at 00:00 UTC.

4.1.7 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, biomass burning 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 (NAVGEM), which is an operational global weather prediction system developed by the United States Navy (Hogan et al., 2014). NAAPS uses BB smoke source from the Fire Locating and Modeling of Burning Emissions (FLAMBE) inventory, which is based on near-real time MODIS fire hotspot data (Reid et al., 2009). Dust emissions for the NAAPS model were generated with the methods documented in Westphal et al. (1988), while the dust scheme algorithms were based on Westphal et al. (2009). This study employed the NAAPS global 3-hourly AOD and surface PM$_{2.5}$ concentrations at a horizontal resolution of $0.333^\circ \times 0.333^\circ$ and initialized each day at 00:00 UTC.
4.2 Multi-Model Ensemble Forecasts

The ensemble forecasts were created using the mean AOD and PM$_{2.5}$ generated by the individual models. All data were interpolated to a unified horizontal grid of 12 km $\times$ 12 km before calculating the ensemble mean values. It should be noted that the individual model and the ensemble simulations were evaluated with the observations grid by grid, and any grids containing missing data were ignored from the calculation.

4.2.1 Ensemble AOD Forecasts

For the 2020 Gigafire events, the ensemble forecasts were reproduced from August to September 2020 using the mean values of AOD generated by seven models: GMU-CMAQ, NACC-CMAQ, HYSPLIT, ICAP-MME, GEFS-Aerosols, GEOS-5, and NAAPS models. The ensemble AOD forecasts were simulated near VIIRS equatorial crossing time (13:30 LST) and mean value of the average of AOD near equatorial crossing times of MODIS Terra (10:30 LST) and Aqua (13:30 LST) as MAIAC processes MODIS Terra and Aqua jointly as a single sensor. The performance of the ensemble AOD mean near VIIRS passing time was evaluated against the VIIRS AOD. While the performance of the ensemble mean of average AOD near MODIS Terra and Aqua passing time was evaluated against MAIAC AOD retrievals.

For the 2021 Spring Dust Season, the ensemble forecasts of AOD were reproduced from January to March 2021 using the mean values of AOD from six models: GMU-CMAQ, NACC-CMAQ, ICAP-MME, GEFS-Aerosols, GEOS-5, and NAAPS models. The ensemble AOD was simulated using the mean value of AOD near VIIRS-
SNPP equatorial crossing time (13:30 LST). The performance of the ensemble AOD mean near VIIRS-SNPP passing time was evaluated against the VIIRS enhanced DT AOD and VIIRS DB AOD.

4.2.2 Ensemble PM$_{2.5}$ Forecasts

The ensemble PM$_{2.5}$ forecasts during the 2020 Gigafire events and the 2021 Spring Dust Season were simulated using the mean values of PM$_{2.5}$ concentrations generated by six models (for wildfire case), and five models (for dust storm case): GMU-CMAQ, NACC-CMAQ, HYSPLIT (only wildfire), GEFS-Aerosols, GEOS-5, and NAAPS models. The performance of ensemble mean in forecasting PM$_{2.5}$ concentrations for both cases was verified by comparing model simulations against daily average PM$_{2.5}$ observations from AirNow with the evaluation time starting from 12:00 UTC to 11:00 UTC of the next day.

4.3 Ensemble Probability of PM$_{2.5}$ Exceedance Forecast

The GMU-CMAQ, NACC-CMAQ, HYSPLIT (only wildfire), GEFS-Aerosol, GEOS-5, 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 numbers of models that forecast PM$_{2.5}$ exceedances (concentrations >35 µg/m$^3$) during the 2020 Gigafire events and the 2021 Spring Dust Season. The probability result ranges from 0% (none of the models forecast the exceedances; very unlikely to occur) to 100% (all models forecast the exceedances; very likely to occur):
Equation 1 Ensemble Probability of Exceedance Forecast

\[ P(A) = \frac{\text{Number of models that forecast the exceedances}}{\text{Total number of models}} \times 100\% \]
CHAPTER 5: EVALUATION METHODOLOGY

5.1 Statistical Metrics

The AOD and surface PM$_{2.5}$ concentrations simulated by the ensemble mean and individual models were evaluated with AOD retrievals from VIIRS enhanced DT, VIIRS DB (for the 2021 Spring Dust Season) and MAIAC (for the 2020 Gigafire events) and observed surface PM$_{2.5}$ from the AirNow ground monitoring network. A suite of statistical metrics, including root mean square error ($RMSE$), correlation ($CORR$), absolute fractional bias ($FB$), mean bias ($MB$), mean error ($ME$), normalized mean bias ($NMB$), and normalized mean error ($NME$) were calculated using the following formulas:

Equation 2 Root Mean Square Error

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

Equation 3 Correlation

$$CORR = \frac{N \sum_{i=0}^{N} O_i M_i - \sum_{i=0}^{N} O_i \sum_{i=0}^{N} M_i}{\sqrt{N \sum_{i=0}^{N} O_i^2 - (\sum_{i=0}^{N} O_i)^2} \sqrt{N \sum_{i=0}^{N} M_i^2 - (\sum_{i=0}^{N} M_i)^2}}$$

Equation 4 Absolute Fractional Bias

$$FB = 2 \times \frac{\sum_{i=0}^{N} |O_i - M_i|}{\sum_{i=0}^{N} |O_i + M_i|}$$
Equation 5 Mean Bias

\[ MB = \frac{1}{N} \sum_{i=0}^{N} (M_i - O_i) \]

Equation 6 Mean Error

\[ ME = \frac{1}{N} \sum_{i=0}^{N} |M_i - O_i| \]

Equation 7 Normalized Mean Bias

\[ NMB = \frac{\sum_{i=0}^{N}(M_i - O_i)}{\sum_{i=0}^{N} O_i} \]

Equation 8 Normalized Mean Error

\[ NME = \frac{\sum_{i=0}^{N}|M_i - O_i|}{\sum_{i=0}^{N} O_i} \]

Where \( M_i \) represents the \( i \)th model forecast, \( O_i \) is the \( i \)th observation, and \( N \) is the total number of observations and time-space matched prediction during the study periods.

5.2 Categorical Metrics

In addition to a traditional suite of statistical metrics, we employed two categorical metrics: the area hit rate (aH), and the area false alarm ratio (aFAR) (Kang et al., 2007) to supplementarily measure the performance of individual models, ensemble mean, and ensemble probability in forecasting PM\(_{2.5}\) exceedances (24-hr PM\(_{2.5}\) concentrations greater than 35 µg/m\(^3\) based on NAAQS). These two metrics were calculated based on pairs of observed and predicted PM\(_{2.5}\) exceedances by considering four possible scenarios: (a) a forecasted exceedance that is not observed; (b) a forecasted
exceedance that is observed; (c) an exceedance that is neither forecasted nor observed; (d) an observed exceedance that is not forecasted (Figure 3). The $aH$ and $aFAR$ values are determined by matching observed and forecasted exceedances within a designated area surrounding the center of the observation location. In the present study, we used an area of $0.5^\circ \times 0.5^\circ$ centered at each AirNow site’s location. The area hit rate $aH$ refers to the number of hits if a forecasted exceedance is observed within the designated area (Eq.9). The $aFAR$ (Eq.10) refers to the false-alarm ratio if a forecasted exceedance is not observed within the designated area:

**Equation 9 area Hit rate**

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

**Equation 10 area False Alarm Ratio**

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

Where $Aa$ is the number of forecasted exceedances that are not observed, $Ab$ is the number of forecasted exceedances that are observed, and $Ad$ is the number of observed exceedances that were not forecasted.
5.3 Overall Rating

The overall rating (RANK) was used to determine the comprehensive forecasting performances of individual models and ensemble mean during the study periods. In the case of PM$_{2.5}$ evaluation, the RANK was derived from the sum of the normalized CORR, $FB$, $aH$, and $aFAR$ (Eq.11). In the case of AOD evaluation, the RANK was calculated using the sum of the normalized CORR and $FB$ (Eq.12). PM$_{2.5}$ RANK ranges from 0 to 4 (from worst to best), while AOD RANK ranges from 0 to 2:
Equation 11 Overall Rating for PM$_{2.5}$

\[
RANK_{PM_{2.5}} = \frac{CORR + 1}{2} + \left(1 - \frac{FB}{2}\right) + \frac{aH}{100\%} + \left(1 - \frac{aFAR}{100\%}\right)
\]

Equation 12 Overall Rating for AOD

\[
RANK_{AOD} = \frac{CORR + 1}{2} + \left(1 - \frac{FB}{2}\right)
\]
CHAPTER 6: EVALUATION OF ENSEMBLE FORECAST OF WILDFIRES

In this chapter, an ensemble mean based on the unweighted arithmetic mean of individual models, including GMU-CMAQ, NACC-CMAQ, GEFS-Aerosols, GEOS-5, HYSPLIT, ICAP-MME, and NAAPS, was created and evaluated with satellite and ground observations during the 2020 Gigafire events (August-September 2020). The forecasting performance of the ensemble mean was also compared with ensemble members to assess whether the ensemble mean can outperform the top performers among these members. Evaluation results of AOD simulations validated against VIIRS enhanced DT AOD and MAIAC AOD retrievals, and the evaluation results of PM$_{2.5}$ simulations compared against AirNow PM$_{2.5}$ observations were analyzed by calculating average statistical metrics and the overall rating (RANK) for AOD and PM$_{2.5}$ simulations over the study period.

6.1 Ensemble Performance in Forecasting AOD

The ensemble mean shows fairly good performance in simulating AOD. For instance, contour maps of AOD forecasts, VIIRS AOD and MAIAC AOD retrievals on August 22$^{nd}$, 2020, in Figures 4a-4g and 4a-4g indicates that the AOD simulations from all the Model-1 to 7 underestimated AOD values over the western U.S. while Model-5 overestimated AOD values primarily in California (Figures 4c, 4e, and Figures 5c, 5e). In
comparison, the ensemble mean AOD simulations slightly overestimated AOD values for the most parts of Northern California and underestimated AOD values predominantly over Montana, Wyoming, Colorado, Nebraska, and Kansas where complex geographic formations, such as the Colorado Plateau-Central Rockies areas, are located (Figures 4h and 5h). However, the majority of areas showing high AOD values in the ensemble forecast match fairly well with the observations.

As shown in Figures 7c and 8c in the time series plots of mean bias ($MB$), the ensemble mean and most individual models slightly underestimated AOD values almost the entire period, especially during the extreme fires in the middle of September 2020 leading to relatively high negative $MB$ values during this time. Similar to previous results, Model-5 overestimated the AOD values with relatively high positive $MB$ values during the same period. From Table 1, the average $MB$ values of the ensemble mean for AOD over the whole period was reduced to -0.104 for the VIIRS AOD case and was greatly reduced to -0.068 for the MAIAC AOD case. Both values are closer to zero relative to most individual models, meaning that the ensemble mean significantly reduces bias and uncertainties in AOD forecasting. In addition, the time series plots of correlation ($CORR$) (Figures 7a and 8a) display less fluctuated correlation lines of the ensemble mean in relation to most individual models, resulting in being the best in correlation for the entire period in the VIIRS and MAIAC AOD cases shown in Table 1.

Considering the overall rating ($RANK$) (Table 1), the ensemble mean scores second in $RANK$ with the $RANK$ values being only 4.8% and 4.0% lower than the top rank model based on VIIRS AOD and MAIAC AOD retrievals, respectively. Although
the averages of \textit{RMSE}, \textit{NMB}, \textit{NME}, \textit{ME}, and \textit{FB} of the ensemble mean did not outperform the first rank individual model, they still rank either second or third place. All the results point to the ensemble mean having a beneficial effect in reducing the bias in AOD forecasting, especially when the wildfires are extremely intense. 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, which are particularly degraded by errors in emission inventory and smoke plume algorithms implemented to each model.

6.2 Ensemble Performance in Forecasting Surface PM$_{2.5}$ Concentration

The ensemble mean show fairly well forecasting performance of surface PM$_{2.5}$ during extreme wildfires, such as the PM$_{2.5}$ forecasts on August 22$^{\text{nd}}$, 2020 (Figure 6). Figures 6a, 6c, 6d, 6e, and 6f show that Model-1, 3, 4, 5, and 7 overestimated PM$_{2.5}$ concentrations largely in the western U.S. and partially in the Central and southern United States. In contrast, the ensemble mean predominantly shows overestimated PM$_{2.5}$ simulations in Northern California (Figure 6g). 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 6h).

The positive mean bias (\textit{MB}) values of the ensemble mean and the individual models in Figure 9c 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 \textit{MB} values (underestimation of PM$_{2.5}$ concentrations). The discrepancies in PM$_{2.5}$ simulations
during the wildfires were also indicated by varying values of root mean square (\textit{RMSE}), mean error (\textit{ME}), normalized mean bias (\textit{NMB}), normalized mean error (\textit{NME}), and absolute fractional bias (\textit{FB}), as shown in Figures 9b, 9d, 9e, 9f, and 9g. Overall, the ensemble mean reduced the positive average mean bias (\textit{MB}) to 7.4 and lowered the absolute fractional bias (\textit{FB}) to 0.54, which is the top rank in \textit{FB}. It also yields the highest average correlation (\textit{CORR}) value of 0.603 (Table 1) due to consistent correlation values for the entire period compared to that of individual models (Figure 9a). The \textit{RMSE}, \textit{NMB}, and \textit{NME} values of the ensemble mean are on average lower than those values of most individual models, as shown in Table 1.

Analysis of the forecasting performance of daily \textit{PM$_{2.5}$} exceedances (concentrations $>$35 $\mu$g/m$^3$) showed the ensemble mean substantially increased the area hit rate (\textit{aH}), particularly in the middle of September when the extremely intense wildfires occurred (Figure 9g). As a result, the ensemble mean achieves the highest average \textit{aH} value of 86.845\% (Table 1). This suggests that the ensemble mean can predict more than 86\% of the observed \textit{PM$_{2.5}$} exceedances during extreme wildfires. Due to relatively high correlation, high \textit{aH}, low \textit{aFAR}, and low \textit{FB} values, the ensemble mean performs highly in \textit{RANK} (2.825). These results suggest that the ensemble forecast has a practical advantage in reducing bias in individual forecasts of \textit{PM$_{2.5}$} and allowing effective probabilistic forecasts of \textit{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 \textit{PM$_{2.5}$} prediction. The model that performs highly in \textit{RANK} for the AOD prediction is different from that of the \textit{PM$_{2.5}$} prediction.
6.3 Overall Performance of Ensemble Wildfire Forecast

Figures 10a-10c show the time series of the overall rating (RANK) for AOD and PM$_{2.5}$ predicted by the ensemble mean and individual models compared against three observation datasets: VIIRS and MAIAC AOD, and AirNow surface PM$_{2.5}$ concentrations, respectively. The ensemble mean shows persistently high RANK values throughout the study period, suggesting that the ensemble forecast overall is more reliable and performs better than most of the members. In addition, it can partially reduce the bias as shown in Table 1 due to the fact that the ensemble mean is calculated by averaging each of the individual model simulation results. Therefore, if most individual models underestimated (negative bias) or overestimated (positive bias) the AOD values and PM$_{2.5}$ concentrations for almost the entire period, the bias values of the ensemble mean become more negative or positive than the top-ranked model with the lowest bias. As a result, the ensemble forecast will show significant and effective improvements in forecasting if there are complementary underestimation and overestimation by individual models.

Underestimation of AOD values and overestimation of PM$_{2.5}$ concentrations in the model simulations may have occurred since the August 2020 complex wildfires that became much more intense during the middle of September 2020, generating very thick smoke cover. The smoke could in turn make the biomass burning emissions applied to each model inaccurate and may generate a large error in smoke inventories. Furthermore, as the fire becomes stronger, the plume injection height gets deeper and creates misrepresented vertical emissions within the planetary boundary layer (PBL) generated
by each individual model. These two factors are considered important sources of uncertainties in air quality forecasts during wildfire events (Carter et al., 2020; Pan et al., 2020b; and Ye et al., 2021). The impact of thick wildfire smoke also challenges the use of satellite AOD retrievals for evaluating ensemble forecasts since retrievals over heavy smoke plumes may be masked as clouds and vertical distributions of smoke are difficult to measure, which affects the accuracy of retrieval AOD products. In addition, a variety of input data sets, such as meteorological fields and chemical transports (F. Li et al., 2019; Y. Li et al., 2020) and plume rise schemes (Briggs, 1969; Freitas et al., 2007; Paugam et al., 2016; Sofiev et al., 2012; Stein et al., 2009; Vernon et al., 2018; Zhu et al., 2018), implemented differently in each model and can also impact the AOD and PM$_{2.5}$ forecasting performance (Delle Monache and Stull, 2003; Kumar et al., 2020).
Figure 4. AOD predicted by seven individual models (a-g) and the ensemble mean (h), compared with VIIRS enhanced Dark Target (DT) AOD retrievals (i) on August 22, 2020 (during the 2020 Gigafire events).
Figure 5. AOD predicted by seven individual models (a-g) and the ensemble mean (h), compared with MAIAC AOD retrievals (i) on August 22, 2020 (during the 2020 Gigafire events).
Figure 6. Surface PM$_{2.5}$ concentrations predicted by six individual models (a-f) and the ensemble mean (g), compared with AirNow PM$_{2.5}$ observations (h) on August 22, 2020 (during the 2020 Gigafire events).
Figure 7. Time series of RMSE (a), \textit{CORR} (b), \textit{MB} (c), \textit{ME} (d), \textit{NMB} (e), \textit{NME} (f), and \textit{FB} (g) of AOD for the 2020 Gigafire events during August-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.
Figure 8. Time series of RMSE (a), CORR (b), MB (c), ME (d), NMB (e), NME (f), and FB (g) of AOD for the 2020 Gigafire events during August-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 the MAIAC AOD retrievals.
Figure 9. Time series of RMSE (a), CORR (b), MB (c), ME (d), NMB (e), NME (f), aH (g), aFAR (h), and FB (i) of PM$_{2.5}$ for the 2020 Gigafire events during August-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.
Figure 10. Time series of the overall rating (\textit{RANK}) for AOD and PM$_{2.5}$ simulated by the ensemble mean and individual models. The \textit{RANK} is calculated with four statistical metrics by comparing model predictions against AOD retrievals from VIIRS (a) and MAIAC (b), and surface PM$_{2.5}$ observations from AirNow (c) during the 2020 Gigafire events (August-September 2020).
Table 1. Overall ensemble mean and individual model performances in forecasting AOD values and PM$_{2.5}$ concentrations during the 2020 Gigafire events (August-September 2020) based on the evaluation of the values of RMSE, CORR, NMB, NME, MB, ME, FB aH and aFAR, and overall rating (RANK). The best results of each statistical metric and RANK are highlighted in bold.

<table>
<thead>
<tr>
<th>Cases</th>
<th>Models</th>
<th>RMSE</th>
<th>CORR</th>
<th>NMB</th>
<th>NME</th>
<th>MB</th>
<th>ME</th>
<th>FB</th>
<th>aH(%)</th>
<th>aFAR(%)</th>
<th>RANK</th>
</tr>
</thead>
<tbody>
<tr>
<td>AOD simulations compared against VIIRS retrievals</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model-1</td>
<td>0.479</td>
<td>0.533</td>
<td>-0.281</td>
<td>0.587</td>
<td>-0.085</td>
<td>0.186</td>
<td>0.682</td>
<td></td>
<td></td>
<td></td>
<td>1.425</td>
</tr>
<tr>
<td>Model-2</td>
<td>0.360</td>
<td>0.477</td>
<td>-0.512</td>
<td>0.626</td>
<td>-0.172</td>
<td>0.199</td>
<td>0.852</td>
<td></td>
<td></td>
<td></td>
<td>1.313</td>
</tr>
<tr>
<td>Model-3</td>
<td>0.551</td>
<td>0.296</td>
<td>-0.565</td>
<td>0.979</td>
<td>-0.153</td>
<td>0.249</td>
<td>1.398</td>
<td></td>
<td></td>
<td></td>
<td>0.949</td>
</tr>
<tr>
<td>Model-4</td>
<td>0.324</td>
<td>0.480</td>
<td>-0.339</td>
<td>0.597</td>
<td>-0.116</td>
<td>0.183</td>
<td>0.726</td>
<td></td>
<td></td>
<td></td>
<td>1.377</td>
</tr>
<tr>
<td>Model-5</td>
<td>1.076</td>
<td>0.429</td>
<td>0.458</td>
<td>1.100</td>
<td>0.201</td>
<td>0.387</td>
<td>0.852</td>
<td></td>
<td></td>
<td></td>
<td>1.289</td>
</tr>
<tr>
<td>Model-6</td>
<td>0.352</td>
<td>0.516</td>
<td>-0.544</td>
<td>0.660</td>
<td>-0.164</td>
<td>0.203</td>
<td>0.931</td>
<td></td>
<td></td>
<td></td>
<td>1.293</td>
</tr>
<tr>
<td>Model-7</td>
<td>0.332</td>
<td>0.506</td>
<td>-0.219</td>
<td>0.540</td>
<td>-0.081</td>
<td>0.167</td>
<td>0.617</td>
<td></td>
<td></td>
<td></td>
<td>1.444</td>
</tr>
<tr>
<td>Ensemble Mean</td>
<td>0.355</td>
<td>0.569</td>
<td>-0.278</td>
<td>0.566</td>
<td>-0.079</td>
<td>0.175</td>
<td>0.664</td>
<td></td>
<td></td>
<td></td>
<td>1.452</td>
</tr>
<tr>
<td>AOD simulations compared against MAIAC retrievals</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model-1</td>
<td>0.385</td>
<td>0.560</td>
<td>-0.241</td>
<td>0.535</td>
<td>-0.052</td>
<td>0.147</td>
<td>0.608</td>
<td></td>
<td></td>
<td></td>
<td>1.476</td>
</tr>
<tr>
<td>Model-2</td>
<td>0.282</td>
<td>0.509</td>
<td>-0.478</td>
<td>0.558</td>
<td>-0.136</td>
<td>0.153</td>
<td>0.741</td>
<td></td>
<td></td>
<td></td>
<td>1.384</td>
</tr>
<tr>
<td>Model-3</td>
<td>0.485</td>
<td>0.351</td>
<td>-0.532</td>
<td>0.963</td>
<td>-0.121</td>
<td>0.215</td>
<td>1.346</td>
<td></td>
<td></td>
<td></td>
<td>1.003</td>
</tr>
<tr>
<td>Model-4</td>
<td>0.263</td>
<td>0.522</td>
<td>-0.280</td>
<td>0.535</td>
<td>-0.077</td>
<td>0.142</td>
<td>0.628</td>
<td></td>
<td></td>
<td></td>
<td>1.447</td>
</tr>
<tr>
<td>Model-5</td>
<td>1.048</td>
<td>0.455</td>
<td>0.587</td>
<td>1.146</td>
<td>0.237</td>
<td>0.375</td>
<td>0.824</td>
<td></td>
<td></td>
<td></td>
<td>1.315</td>
</tr>
<tr>
<td>Model-6</td>
<td>0.289</td>
<td>0.543</td>
<td>-0.497</td>
<td>0.618</td>
<td>-0.124</td>
<td>0.163</td>
<td>0.847</td>
<td></td>
<td></td>
<td></td>
<td>1.348</td>
</tr>
<tr>
<td>Model-7</td>
<td>0.259</td>
<td>0.564</td>
<td>-0.152</td>
<td>0.467</td>
<td>-0.043</td>
<td>0.127</td>
<td>0.514</td>
<td></td>
<td></td>
<td></td>
<td>1.525</td>
</tr>
<tr>
<td>Ensemble Mean</td>
<td>0.303</td>
<td>0.615</td>
<td>-0.218</td>
<td>0.509</td>
<td>-0.042</td>
<td>0.140</td>
<td>0.577</td>
<td></td>
<td></td>
<td></td>
<td>1.521</td>
</tr>
<tr>
<td>PM$_{2.5}$ simulations compared against AirNow observations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model-1</td>
<td>25.612</td>
<td>0.543</td>
<td><strong>0.177</strong></td>
<td>0.618</td>
<td>3.373</td>
<td>9.971</td>
<td>0.545</td>
<td>69.080</td>
<td>44.294</td>
<td></td>
<td>2.814</td>
</tr>
<tr>
<td>Model-2</td>
<td><strong>16.960</strong></td>
<td>0.490</td>
<td>-0.262</td>
<td><strong>0.509</strong></td>
<td>-4.885</td>
<td><strong>8.022</strong></td>
<td>0.592</td>
<td>39.674</td>
<td>23.737</td>
<td></td>
<td>2.735</td>
</tr>
<tr>
<td>Model-3</td>
<td>20.571</td>
<td>0.433</td>
<td>-0.483</td>
<td>0.962</td>
<td>-4.654</td>
<td>10.645</td>
<td>1.321</td>
<td>71.714</td>
<td>45.416</td>
<td></td>
<td>2.401</td>
</tr>
<tr>
<td>Model-4</td>
<td>51.955</td>
<td>0.492</td>
<td>1.369</td>
<td>1.567</td>
<td>19.804</td>
<td>23.004</td>
<td>0.892</td>
<td>79.159</td>
<td>75.574</td>
<td></td>
<td>2.382</td>
</tr>
<tr>
<td>Model-5</td>
<td>51.940</td>
<td>0.436</td>
<td>0.748</td>
<td>1.122</td>
<td>13.020</td>
<td>18.421</td>
<td>0.772</td>
<td>80.995</td>
<td>68.684</td>
<td></td>
<td>2.497</td>
</tr>
<tr>
<td>Model-7</td>
<td>39.984</td>
<td>0.512</td>
<td>0.814</td>
<td>1.008</td>
<td>12.296</td>
<td>15.912</td>
<td>0.697</td>
<td>80.789</td>
<td>62.514</td>
<td></td>
<td>2.632</td>
</tr>
<tr>
<td>Ensemble Mean</td>
<td>25.925</td>
<td><strong>0.603</strong></td>
<td>0.442</td>
<td>0.685</td>
<td>7.403</td>
<td>11.161</td>
<td><strong>0.537</strong></td>
<td><strong>86.845</strong></td>
<td>60.515</td>
<td></td>
<td><strong>2.825</strong></td>
</tr>
</tbody>
</table>

### 6.4 Ensemble Probability Forecast of PM$_{2.5}$ Exceedances

In general, the ensemble probability shows fairly good performance in forecasting PM$_{2.5}$ exceedances during the 2020 Gigafire events. Figure 11 depicts a contour map of
ensemble probability forecast values overlaid by the actual exceedance (binary) over the AirNow sites across the CONUS. The probability ranges from 16.67% (exceedances predicted by only one model; very unlikely to occur) to 100% (exceedances predicted by all six models; very likely to occur). The more models that forecast the exceedance for each grid, the higher probability that the exceedances will occur in that grid. As shown in Figure 11, the contours of high ensemble probability values of 83.33% (five models; orange) and 100% (all six models; red) were displayed mainly in California, which collocated well with the AirNow exceedance measurements (marked as filled red circles). However, the AirNow observed exceedances in the downwind region (Idaho and Montana) were only captured by four of six models, giving a probability forecast of 66.67% (four of six models; 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 validated their performance in 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 shown as time series plots of $aH$ and $aFAR$ in Figures 12a and 12b. The average $aH$ and $aFAR$ values are listed in Table 2. High $aH$ value and low $aFAR$ values suggest good agreement between model simulations and observations. As displayed in the time series plots of $aH$ and $aFAR$ (Figure 12a and 12b) and the average $aH$ and $aFAR$ values (Table 2), the lowest ensemble probability of 16.67% shows constantly high $aH$ and high $aFAR$ throughout the study period, resulting in being the greatest $aH$ value of 93.985 (top-ranked) and also the
highest $aFAR$ value of 78.003 (lowest-ranked) on average, while the highest ensemble probability of 100% show persistently and relatively low $aH$ and low $aFAR$ all the time, resulting in holding the lowest average $aH$ value of 14.725 (lowest-ranked) and the lowest $aFAR$ value of 1.537 (top-ranked).

The evaluation results imply that including a small number of models in the ensemble or the low ensemble probability shows better performance in forecasting observed exceedances across the CONUS because some exceedances predicted by any individual model or the lowest ensemble probability were true exceedances associated with wildfires, especially in the wildfire active regions, which frequently matched the AirNow observations, resulting in high $aH$. 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 $aFAR$. 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 due to the fact that their predicted exceedances have been calibrated and verified with the co-existed exceedances predicted by the other participant models included in the ensemble. As a consequence, the areas showing the false exceedances have been reduced or removed, resulting in lower $aH$ and $aFAR$ values.
Figure 11. Ensemble probability forecast of PM$_{2.5}$ exceedances on August 22nd, 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; unlikely to occur) (light blue) to 100% (all six models forecast the PM$_{2.5}$ exceedances; very likely to occur) (red). The PM$_{2.5}$ exceedances observed by the AirNow sites are displayed in the red/green circles (red means an exceedance recorded by the monitor, and green means no exceedance recorded).
Figure 12. Time series plots of $aH$ (a) and $aFAR$ (b) values during the 2020 Gigafire events (August-September 2020) for the ensemble probability of PM$_{2.5}$ exceedance forecast. Ensemble probability values range from 16.67% (one out of six models) to 100% (all six models).

Table 2. Averaged $aH$ and $aFAR$ values of ensemble probability of PM$_{2.5}$ exceedance forecast during the 2020 Gigafire events (August-September 2020), comparing between simulated PM$_{2.5}$ exceedances and observed PM$_{2.5}$ exceedances obtained from AirNow.

<table>
<thead>
<tr>
<th>Ensemble Probability</th>
<th>Statistical Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$aH$</td>
</tr>
<tr>
<td>16.67%</td>
<td>93.985</td>
</tr>
<tr>
<td>33.33%</td>
<td>88.398</td>
</tr>
<tr>
<td>50%</td>
<td>79.305</td>
</tr>
<tr>
<td>66.67%</td>
<td>69.716</td>
</tr>
<tr>
<td>83.33%</td>
<td>48.104</td>
</tr>
<tr>
<td>100%</td>
<td>14.725</td>
</tr>
</tbody>
</table>

$aH$: area hit rate; $aFAR$: area false alarm ratio
CHAPTER 7: EVALUATION OF ENSEMBLE FORECAST OF DUST STORMS

In this chapter, an ensemble AOD and PM$_{2.5}$ forecasts based on the unweighted arithmetic mean of six individual models, including GMU-CMAQ, NACC-CMAQ, GEFS-Aerosols, GEOS-5, ICAP-MME, and NAAPS, were created and evaluated with VIIRS enhanced DT AOD and VIIRS DB AOD, and AirNow surface PM$_{2.5}$ observations, respectively, during the 2021 Spring Dust Season (January-March 2021). The forecasting performance of the ensemble mean was also intercompared with the ensemble members to assess whether the ensemble mean can outperform the top performers among these members. The evaluation results were analyzed by calculating average statistical metrics and the overall rating ($RANK$) over the study period.

7.1 Ensemble AOD Forecasting Performance on March 16$^{th}$, 2021

The AOD simulations were initially evaluated against VIIRS enhanced Dark Target (DT) over dark and bright surfaces AOD on March 16$^{th}$, 2021. As a consequence of limitations in the AOD detectability of the enhanced DT algorithm, whose maximum retrieved AOD did not pass the data quality test at initial data processes, we validated the AOD simulations with the VIIRS Deep Blue (DB) AOD retrievals.

In general, the ensemble mean shows slightly improved performance in predicting observed dust AOD as demonstrated in contour maps of AOD forecasts, VIIRS enhanced
DT AOD observations, and VIIRS DB AOD observations on March 16\textsuperscript{th}, 2021 (Figures 13a-13i). As illustrated in the contour map of VIIRS DB AOD (Figure 13i), the observed high AOD associated with the dust storm almost blanketed entire areas of the Chihuahuan Desert, especially in El Paso, and partially covered the downwind areas in southern New Mexico. Unfortunately, due to poor quality control, the high AOD retrieved by VIIRS enhanced DT over the Chihuahuan Desert was missing, as shown in Figure 13h. Model-1, Model-2, and Model-4 were able to simulate dust storms primarily in west Texas and incompletely over the downwind areas in southern New Mexico. However, compared to VIIRS DB AOD, these models still underestimated AOD over the active dust areas (Figures 13a, 13b, 13d). Meanwhile, Model-3, Model-5, and Model-6 were not able to simulate dust storms over the Chihuahuan Desert in parts of southern New Mexico, western Texas, and along the Mexico-United States border, where the powerful dust storm originated from and blew through (Figures 13c, 13e, 13f). Considering background AOD across the CONUS, all individual models were likely to simulate fairly higher background AOD compared to the satellite observations. In addition, all models except Model-5 predicted relatively high AOD elsewhere, primarily in the southeastern coast of the United States and partially in the northeastern United States. However, we were unable to compare this predicted high AOD with the VIIRS enhanced DT and DB AOD observations due to missing retrieved AOD data therein. These predicted high AOD may be model overestimations due to other emissions like prescribed fires or anthropogenic emissions, while underpredicted AOD may be a consequence of large uncertainties in dust emissions, transports, and depositions, as well
as variability in dust parameterizations in each model. The individual models underestimated AOD near the dust source regions in western Texas and southern New Mexico and overestimated AOD in the Southeast Coast of the U.S., causing the ensemble mean to demonstrate these same patterns, as shown in Figure 13g.

The performance of the ensemble mean in AOD forecasting on March 16th, 2021 was analyzed through the statistical metrics, as shown in Table 4. Based on VIIRS enhanced DT, the ensemble mean shows the absolute fractional bias ($FB$) reduced to 0.625 at the lowest value from a range of 0.643 to 1.440, lowered mean bias ($MB$; 0.046), improved correlation ($CORR$; 0.188), and had the second highest overall rating ($RANK$; 1.281). In comparison with the VIIRS DB, the ensemble mean demonstrated declining $MB$ (0.094), the second highest correlation (0.397), and the second highest $RANK$ (1.174) among six members. This result also points to the fact that the underestimations of AOD in the Chihuahuan Desert (in western Texas and southern New Mexico) did not always translate significantly to the total biases in AOD forecasting across the CONUS due to small-scaled dust affected areas. In this case, the positive biases may be due to the effects of high background AOD simulations. In addition, the inconsistency in model biases between the evaluations with VIIRS enhanced DT and VIIRS DB is a result of AOD being retrieved differently by DT and DB algorithms.

Due to the uncertainties in the AOD simulations over the CONUS, we also generated the ensemble AOD forecast for the local active dust region domain covering the Chihuahuan Desert in Mexico across western Texas and southern New Mexico ($25^\circ$N to $40^\circ$N, -$110^\circ$W to -$95^\circ$W) (Figure 14g). The performance of the ensemble mean in
predicting AOD over this constrained domain was compared with the individual models and then verified with the VIIRS DB AOD by calculating the correlation between predicted AOD and observed AOD on March 16th, 2021. Table 3 shows the comparison of the correlation between observed AOD and predicted AOD over the CONUS domain and the active dust region domain on March 16th, 2021. For the CONUS domain, Model-1 yields the highest correlation (0.451), followed by the ensemble mean scoring the second highest in correlation (0.397). In the case of the active dust regions domain, the correlation of Model-1, whose predicted dust related AOD matched VIIRS DB AOD the most (Figure 14a), improved and became the top rank in correlation (0.454). Whereas the correlations of Model-3 and Model-6, which failed to simulate dust storms in the Chihuahuan Desert (Figures 14c and 14f), were reduced to negative values. Consistent with the large underestimations of dust related AOD by most individual models, the correlation of the ensemble mean was also reduced, but by only 1% (0.392). This suggests an effective capability of the ensemble approach to reduce the biases in model forecasts over the active dust regions. Overall, these results revealed the actual forecasting performance of individual models and the ensemble mean in local dust source areas and also emphasized the great impact of the uncertainties in the model simulations on the accuracy of dust AOD forecasts over the CONUS.
Figure 13. AOD predicted by six individual models (a-f) and the ensemble mean (g), compared with VIIRS enhanced Dark Target (DT) AOD retrievals (h) and VIIRS Deep Blue (DB) AOD retrievals (i) on March 16th, 2021 (during the 2021 Spring Dust Season).
Figure 14. AOD predicted by six individual models (a-f) and the ensemble mean (g), compared with VIIRS Deep Blue (DB) AOD retrievals (h) near the active dust regions in western Texas and southern New Mexico on March 16th, 2021 (during the 2021 Spring Dust Season).
Table 3. Correlation between AOD simulations and VIIRS Deep Blue (DB) AOD observations over the Contiguous United States (CONUS) and Active Dust Regions in western Texas and southern New Mexico on March 16th, 2021. The highest correlation is highlighted in bold.

<table>
<thead>
<tr>
<th>Model</th>
<th>CONUS</th>
<th>Active Dust Regions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model-1</td>
<td>0.451</td>
<td>0.454</td>
</tr>
<tr>
<td>Model-2</td>
<td>0.280</td>
<td>0.265</td>
</tr>
<tr>
<td>Model-3</td>
<td>-0.146</td>
<td>-0.323</td>
</tr>
<tr>
<td>Model-4</td>
<td>0.351</td>
<td>0.329</td>
</tr>
<tr>
<td>Model-5</td>
<td>0.081</td>
<td>0.075</td>
</tr>
<tr>
<td>Model-6</td>
<td>0.044</td>
<td>-0.022</td>
</tr>
<tr>
<td>Ensemble Mean</td>
<td>0.397</td>
<td>0.392</td>
</tr>
</tbody>
</table>
Table 4. Overall ensemble mean and individual model performances in forecasting AOD values and PM$_{2.5}$ concentrations on March 16$^{th}$, 2021, based on the evaluation of the values of RMSE, CORR, NMB, NME, MB, ME, FB, aH, and aFAR, and overall rating (RANK). The best results of each statistical metric and RANK are highlighted in bold.

<table>
<thead>
<tr>
<th>Cases</th>
<th>Models</th>
<th>RMSE</th>
<th>CORR</th>
<th>NMB</th>
<th>NME</th>
<th>MB</th>
<th>ME</th>
<th>FB</th>
<th>aH(%)</th>
<th>aFAR(%)</th>
<th>RANK</th>
</tr>
</thead>
<tbody>
<tr>
<td>AOD simulations compared against VIIRS enhanced Dark Target retrievals</td>
<td>Model-1</td>
<td>0.197</td>
<td>0.080</td>
<td>-0.432</td>
<td>0.787</td>
<td>-0.064</td>
<td>0.117</td>
<td>1.004</td>
<td></td>
<td></td>
<td>1.038</td>
</tr>
<tr>
<td></td>
<td>Model-2</td>
<td>0.245</td>
<td>0.094</td>
<td>1.109</td>
<td>1.285</td>
<td>0.164</td>
<td>0.190</td>
<td>0.827</td>
<td></td>
<td></td>
<td>1.134</td>
</tr>
<tr>
<td></td>
<td>Model-3</td>
<td><strong>0.115</strong></td>
<td>0.131</td>
<td><strong>-0.252</strong></td>
<td>0.562</td>
<td><strong>-0.036</strong></td>
<td><strong>0.079</strong></td>
<td>0.643</td>
<td></td>
<td></td>
<td>1.119</td>
</tr>
<tr>
<td></td>
<td>Model-4</td>
<td>0.413</td>
<td>0.276</td>
<td>2.052</td>
<td>2.138</td>
<td>0.304</td>
<td>0.316</td>
<td>1.055</td>
<td></td>
<td></td>
<td>1.110</td>
</tr>
<tr>
<td></td>
<td>Model-5</td>
<td>0.167</td>
<td>-0.113</td>
<td>-0.829</td>
<td>0.844</td>
<td>-0.121</td>
<td>0.124</td>
<td>1.440</td>
<td></td>
<td></td>
<td>0.723</td>
</tr>
<tr>
<td></td>
<td>Model-6</td>
<td>0.123</td>
<td><strong>0.325</strong></td>
<td>-0.393</td>
<td><strong>0.553</strong></td>
<td>-0.058</td>
<td>0.082</td>
<td>0.689</td>
<td></td>
<td></td>
<td><strong>1.318</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Ensemble Mean</strong></td>
<td>0.154</td>
<td>0.188</td>
<td>0.313</td>
<td>0.723</td>
<td>0.046</td>
<td>0.106</td>
<td><strong>0.625</strong></td>
<td></td>
<td></td>
<td><strong>1.281</strong></td>
</tr>
<tr>
<td>AOD simulations compared against VIIRS Deep Blue retrievals</td>
<td>Model-1</td>
<td>0.216</td>
<td><strong>0.451</strong></td>
<td>0.057</td>
<td>0.897</td>
<td><strong>0.005</strong></td>
<td>0.081</td>
<td>0.872</td>
<td></td>
<td></td>
<td><strong>1.290</strong></td>
</tr>
<tr>
<td></td>
<td>Model-2</td>
<td>0.303</td>
<td>0.280</td>
<td>2.355</td>
<td>2.787</td>
<td>0.212</td>
<td>0.251</td>
<td>1.280</td>
<td></td>
<td></td>
<td><strong>1.000</strong></td>
</tr>
<tr>
<td></td>
<td>Model-3</td>
<td>0.252</td>
<td>-0.146</td>
<td>0.279</td>
<td>1.330</td>
<td>0.026</td>
<td>0.126</td>
<td>1.167</td>
<td></td>
<td></td>
<td><strong>0.844</strong></td>
</tr>
<tr>
<td></td>
<td>Model-4</td>
<td>0.438</td>
<td>0.351</td>
<td>3.310</td>
<td>3.675</td>
<td>0.305</td>
<td>0.339</td>
<td>1.384</td>
<td></td>
<td></td>
<td><strong>0.983</strong></td>
</tr>
<tr>
<td></td>
<td>Model-5</td>
<td>0.211</td>
<td>0.081</td>
<td>-0.677</td>
<td><strong>0.819</strong></td>
<td>-0.060</td>
<td><strong>0.073</strong></td>
<td>1.239</td>
<td></td>
<td></td>
<td><strong>0.921</strong></td>
</tr>
<tr>
<td></td>
<td>Model-6</td>
<td><strong>0.205</strong></td>
<td>0.044</td>
<td>-0.151</td>
<td>0.847</td>
<td>-0.013</td>
<td>0.076</td>
<td>0.916</td>
<td></td>
<td></td>
<td><strong>1.064</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Ensemble Mean</strong></td>
<td>0.219</td>
<td>0.397</td>
<td>1.016</td>
<td>1.582</td>
<td>0.094</td>
<td>0.146</td>
<td>1.049</td>
<td></td>
<td></td>
<td><strong>1.174</strong></td>
</tr>
<tr>
<td>PM$_{2.5}$ simulations compared against AirNow observations</td>
<td>Model-1</td>
<td>8.439</td>
<td>0.300</td>
<td>0.092</td>
<td>0.562</td>
<td>0.869</td>
<td>5.289</td>
<td>0.537</td>
<td>20.000</td>
<td>90.090</td>
<td>1.672</td>
</tr>
<tr>
<td></td>
<td>Model-2</td>
<td>7.117</td>
<td>0.242</td>
<td>-0.134</td>
<td><strong>0.478</strong></td>
<td>-1.258</td>
<td>4.501</td>
<td>0.512</td>
<td>0.000</td>
<td>100.000</td>
<td>1.365</td>
</tr>
<tr>
<td></td>
<td>Model-3</td>
<td>11.832</td>
<td>0.122</td>
<td>0.383</td>
<td>0.752</td>
<td>3.620</td>
<td>7.107</td>
<td>0.631</td>
<td>0.000</td>
<td>100.000</td>
<td>1.245</td>
</tr>
<tr>
<td></td>
<td>Model-4</td>
<td>8.656</td>
<td><strong>0.317</strong></td>
<td>0.183</td>
<td>0.623</td>
<td>1.720</td>
<td>5.866</td>
<td>0.571</td>
<td><strong>60.000</strong></td>
<td><strong>70.000</strong></td>
<td><strong>2.273</strong></td>
</tr>
<tr>
<td></td>
<td>Model-6</td>
<td>6.972</td>
<td>0.205</td>
<td><strong>0.075</strong></td>
<td>0.489</td>
<td><strong>0.691</strong></td>
<td>4.523</td>
<td>0.471</td>
<td>0.000</td>
<td>100.000</td>
<td>1.367</td>
</tr>
<tr>
<td></td>
<td><strong>Ensemble Mean</strong></td>
<td><strong>6.921</strong></td>
<td>0.284</td>
<td>0.115</td>
<td>0.478</td>
<td>1.084</td>
<td><strong>4.499</strong></td>
<td><strong>0.452</strong></td>
<td>0.000</td>
<td>100.000</td>
<td>1.416</td>
</tr>
</tbody>
</table>

7.2 Ensemble PM$_{2.5}$ Forecasting Performance on March 16$^{th}$, 2021

Next, the PM$_{2.5}$ simulations were evaluated against AirNow ground observations on March 16$^{th}$, 2021. Overall, the ensemble mean underperformed in predicting dust storms (high PM$_{2.5}$ concentrations) in the Chihuahuan Desert. As shown in a map of AirNow PM$_{2.5}$ observations (Figure 15g), PM$_{2.5}$ concentrations above the National
Ambient Air Quality Standard (NAAQS) (>35 µg/m³) were only observed by three AirNow monitoring sites in western Texas, in El Paso (56 and 55 µg/m³) and Socorro (54 µg/m³), and two sites in southern New Mexico, in Hobbs (40 µg/m³) and Santa Teresa (57 µg/m³) (marked as red filled circles). Figures 15a, 15c, and 15d show that Model-1, Model-3, and Model-4 were able to predict elevated PM$_{2.5}$ concentrations affected by dust storms originating in the Chihuahuan Desert across western Texas and southern New Mexico. However, Model-1 overestimated PM$_{2.5}$ over the dust affected regions, primarily in western Texas and across southwestern New Mexico (>80 µg/m³) (Figure 15a). Model-3 greatly overestimated PM$_{2.5}$ in northwestern Texas (>70 µg/m³), eastern New Mexico (>100 µg/m³), and along the southern Colorado-Kansas border (>100 µg/m³) (Figure 15c), while Model-4 overestimated PM$_{2.5}$ predominantly in southeastern New Mexico-Texas border (>90 µg/m³) (Figure 15d). Model-2 and Model-6 were unable to predict the dust storms in the Chihuahuan Desert, resulting in low simulated PM$_{2.5}$ over western Texas and southern New Mexico (Figures 15b and 15f). Furthermore, elevated PM$_{2.5}$ concentrations were predicted specifically over the southeastern coast of the U.S. by Model-3 (35-75 µg/m³) and Model-4 (35-100 µg/m³), were slightly overestimated in the northeastern U.S. by Model-1 and Model-6 (35-50 µg/m³), but there were no PM$_{2.5}$ concentrations above 35 µg/m³ observed by AirNow sites in these two regions. Despite the underestimation of PM$_{2.5}$ over local active dust regions and significant overestimation of PM$_{2.5}$ in the southeastern coast of the U.S. and the northeastern U.S. from participating models, the ensemble PM$_{2.5}$ forecast underpredicted PM$_{2.5}$ over the Chihuahuan Desert in western Texas and southern New Mexico (10-40 µg/m³) and overpredicted PM$_{2.5}$ in more
constrained areas in western New Mexico (>90 µg/m^3) and along the southern Colorado-Kansas border (>100 µg/m^3), and slightly overpredicted PM$_{2.5}$ in the southeastern coast of the U.S. (35-55 µg/m^3) (Figure 15f). Similar to the AOD forecasts, the predicted elevated PM$_{2.5}$ in the U.S. Southeast and Northeast may be a consequence of prescribed fire and anthropogenic emissions, and the underestimations of PM$_{2.5}$ may arise from errors in the model simulations that are influenced by large uncertainties in dust emissions, variability in dust physical and chemical processes and different dust schemes implemented in each model.

In general, the statistical results in Table 4 indicated the ability of the ensemble mean to moderately reduce biases in model forecasting and improve the accuracy of surface PM$_{2.5}$ forecasts during dust storm events. For instance, the ensemble shows the best RMSE (6.921 from a range of 6.972-11.832) and $FB$ (0.452 from a range of 0.471-0.631), decreased $MB$ (6.921), slightly increased correlation (0.284), and the third highest $RANK$ (1.416) among five members. However, the area hit rate ($aH$) value of 0 and area false alarm ratio ($aFAR$) value of 100 achieved by the ensemble mean suggest that there is no PM$_{2.5}$ exceedance (>35 µg/m$^3$) being observed by any AirNow sites on March 16$^{th}$, 2021 was predicted by the ensemble mean. Therefore, all ensemble predicted PM$_{2.5}$ exceedances were false exceedances. Furthermore, this evaluation result also implied that the underpredictions of PM$_{2.5}$ in the small-scaled local dust associated areas did not significantly contribute to overall biases in PM$_{2.5}$ forecasts on March 16$^{th}$, 2021, whereas the major contribution causing overall positive biases in forecasts was the PM$_{2.5}$ overpredictions over larger areas in the U.S. Northeast and the Southeast.
Figure 15. Surface PM$_{2.5}$ concentrations predicted by five individual models (a-e) and the ensemble mean (f), compared with AirNow PM$_{2.5}$ observations (g) on March 16th, 2021 (during the 2021 Spring Dust Season).
7.3 Ensemble Forecasting Performance during the 2021 Spring Dust Season

To assess the overall performance of the ensemble mean in forecasting AOD and PM$_{2.5}$ during dust storm events, we also conducted an evaluation for the 2021 Spring Dust Season from January to March 2021.

For AOD forecasts, overall, all individual models and the ensemble mean show high positive mean bias ($MB$) during the intense dust storm in the middle of March (Figures 16-17c). The errors in model simulations of AOD throughout the study period were also demonstrated by high varying values of root mean square ($RMSE$), mean error ($ME$), normalized mean bias ($NMB$), normalized mean error ($NME$), and absolute fractional bias ($FB$), as shown in Figures 16-17b, d, e, f, and 16-17g. From Table 5, the average $MB$ of the ensemble mean over the study period was lowered to 0.013 and 0.068 when compared to VIIRS enhanced DT AOD and VIIRS DB AOD, respectively. According to the time series of correlation shown in Figures 16-17a, the ensemble mean correlation was fairly low and fluctuated throughout the study period, indicating a fairly high level of inconsistency in the forecasts. However, compared to the individual models, the correlation of the ensemble mean was slightly stronger, resulting in increased values of 0.093 and 0.323 (the best correlation) based on VIIRS enhanced DT and VIIRS DB, respectively. Regarding the average overall rating ($RANK$) in Table 5, the ensemble mean yields the third best $RANK$ at 1.077 and 1.286, which is 9% and 5% lower than the top rank model, based on VIIRS enhanced DT AOD and VIIRS DB AOD, respectively.

In the forecasts of surface PM$_{2.5}$ concentration, most individual models and the ensemble mean demonstrate the high positive $MB$ values during the intense dust storm
period (the middle of March) (Figure 18c). The errors in model simulations during the
dust storms were indicated by high RMSE, ME, NMB, NME, and FB in the middle of
March, as shown in Figures 18b, d, e, f, and 18g. Table 5 shows that the average MB of
the ensemble mean during January-March 2021 was slightly reduced to 3.139. As shown
in the time series of correlation (Figure 18a), the correlation of the ensemble mean
fluctuated during the intense dust storm period (mid-March), implying a fairly high level
of discrepancies between the ensemble forecasts and the observations. However, the
correlation of the ensemble mean was slightly more consistent and closer to 1 compared
to most of the participating models, resulting in the second best average correlation at
0.395 (Table 5). In addition, the ensemble performance in forecasting the NAAQS PM$_{2.5}$
exceedance (concentrations $>$35 µg/m$^3$) was determined by calculating area hit rate ($aH$)
and area false alarm ratio ($aFAR$). The time series of $aH$ and $aFAR$ (Figures 18g and 18h)
shows that the ensemble mean slightly increased areas hit rate ($aH$) and reduced area
false alarm ratio ($aFAR$), specifically in the middle of March when some individual
models failed greatly to predict the exceedances, characterized by relatively low $aH$ and
high $aFAR$. It should be noted that the concurrent disappearance of any $aH$ and $aFAR$
lines in the time series (Figure 18g) indicates no exceedances observed by any AirNow
sites at that time. Overall, the ensemble mean slightly increased the average $aH$ to
14.822%, meaning that merely 14% of the predicted exceedances were truly observed. It
also slightly reduced the fairly high average $aFAR$ to 62.425%, meaning that about 62%
of the predicted exceedances were not observed (false exceedances) (Table 5). These
results suggested the capability of the ensemble mean to slightly improve the accuracy of
PM$_{2.5}$ exceedance forecasts during dust storms. With fairly weak correlation, low $aH$, and moderately high $aFAR$ and $FB$, the ensemble mean achieves the third highest $RANK$ (2.239), which is 11% lower than the first ranked individual model.

Ultimately, all the results point to the fact that the ensemble mean has the potential to reduce biases in individual PM$_{2.5}$ forecasts and provide an improved probabilistic forecast of AOD and PM$_{2.5}$ during dust storm events. However, its performance is less effective than the wildfire ensemble forecasts.

7.4 Overall Performance of Ensemble Dust Storm Forecasts

Figures 19a-19c show the time series of the overall rating ($RANK$) for AOD and PM$_{2.5}$ predicted by the ensemble mean and participating models during the 2021 Spring Dust Season compared with VIIRS enhanced Dark Target (DT) over dark and bright surfaces AOD, VIIRS Deep Blue (DB) AOD retrievals, and AirNow surface PM$_{2.5}$ observations.

For AOD, the ensemble mean consistently fluctuated at a poor $RANK$ during the intense dust storms in the middle of March for both comparisons with VIIRS enhanced DT and DB (Figures 19a and 19b). This suggests that the ensemble mean can produce moderately reliable forecasts of AOD during the dust storm events. The inconsistent $RANK$ may result from individual model simulation errors caused predominantly by uncertainties in the inputs and model representations of chemical and physical processes. In addition to the systematic errors in the models, the insufficient satellite-retrieved AOD data from limited aerosol detectability over high surface reflectivity areas like deserts and
cloud contamination areas can cause discrepancies in evaluations affecting the overall RANK because the poor maximum AOD retrievals from these areas are occasionally removed by data quality assurance processes (shown as gaps in Figures 13h and 13i) (Hsu et al., 2019; Levy et al., 2015; Sawyer et al., 2020; Zhang et al., 2018).

For surface PM$_{2.5}$, the ensemble mean also fluctuates in RANK over the period (Figure 19c). This fluctuation may be a consequence of model simulation errors and inconsistencies in evaluations due to the scarcity of ground observations in active dust areas near the Chihuahuan Desert in western Texas and southern New Mexico (Figure 15g). Therefore, the ensemble mean overall underperformed in PM$_{2.5}$ forecasting and reproduced moderately reliable PM$_{2.5}$ forecasts during the dust storm events in the Chihuahuan Desert.

In essence, the ensemble forecast improved the predictability of dust related AOD and PM$_{2.5}$ during the 2021 Spring Dust Season in the Chihuahuan Desert, but less successfully than expected due to significant biases in participating model simulations. These biases in the dust air pollution forecasts are frequently a result of the model simulations being affected by large uncertainties in dust and other prescribed emissions (wildfire and anthropogenic emissions), dust circulation and deposition processes, heterogeneity of soil surface properties, and different applications of dust parameterizations, as well as meteorological fields controlling the synoptic scale to mesoscale wind speed and direction. In addition, local dust events in the western U.S. tend to be limited in duration, caused by small-scale wind circulation, and restricted to a particular region, which can all create dust modeling issues.
Furthermore, all statistical analyses suggested that the predictability of ensemble forecasts during the dust storms will be significantly improved if participating models have complementary underestimation and overestimation. Different models have their individual strengths and weaknesses. Although a single model can be excellent at predicting dust AOD, it is not necessarily translated into good surface PM$_{2.5}$ prediction. Our results also demonstrate that the model that performs the best in $RANK$ for AOD prediction is different from the model that is best at PM$_{2.5}$ prediction.
Figure 16. Time series of RMSE (a), CORR (b), MB (c), ME (d), NMB (e), NME (f), and FB (g) of AOD during the 2021 Spring Dust Season from January to March 2021. The AOD simulations by the ensemble mean (solid black line) and individual models (dash lines): Model-1 (blue), Model-2 (light blue), Model-3 (green), Model-4 (pink), Model-5 (orange), and Model-6 (yellow) were compared against VIIRS enhanced Dark Target (DT) AOD retrievals.
Figure 17. Time series of RMSE (a), CORR (b), MB (c), ME (d), NMB (e), NME (f), and FB (g) of AOD during the 2021 Spring Dust Season from January to March 2021. The AOD simulations by the ensemble mean (solid black line) and individual models (dash lines): Model-1 (blue), Model-2 (light blue), Model-3 (green), Model-4 (pink), Model-5 (orange), and Model-6 (yellow) were compared against VIIRS Deep Blue (DB) AOD retrievals.
Figure 18. Time series of RMSE (a), CORR (b), MB (c), ME (d), NMB (e), NME (f), aH (g), aFAR (h), and FB (i) of PM$_{2.5}$ for the 2021 Spring Dust Season during January-March 2021. The PM$_{2.5}$ simulations by the ensemble mean (black solid line) and individual Model-1 (blue), Model-2 (light blue), Model-3 (green), Model-4 (pink), and Model-6 (yellow) were compared against AirNow PM$_{2.5}$ observations.
Figure 19. Time series of the overall rating ($RANK$) for 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 VIIRS enhanced Dark Target (DT) AOD (a) VIIRS Deep Blue (DB) AOD (b) retrievals, and surface PM$_{2.5}$ observations from AirNow (c) during the 2021 Spring Dust Season.
Table 5. Overall ensemble mean and individual model performances in forecasting AOD values and PM$_{2.5}$ concentrations during the 2021 Spring Dust Season (January-March 2021) based on the evaluation of the values of RMSE, CORR, NMB, NME, ME, MB, FB, aH, and aFAR, and overall rating (RANK). The best results of each statistical metric and RANK are highlighted in bold.

<table>
<thead>
<tr>
<th>Cases</th>
<th>Models</th>
<th>RMSE</th>
<th>CORR</th>
<th>NMB</th>
<th>NME</th>
<th>MB</th>
<th>ME</th>
<th>FB</th>
<th>aH(%)</th>
<th>aFAR(%)</th>
<th>RANK</th>
</tr>
</thead>
<tbody>
<tr>
<td>AOD simulations compared against VIIRS Dark Target retrievals</td>
<td>Model-1</td>
<td>0.118</td>
<td>0.083</td>
<td>-0.433</td>
<td>0.746</td>
<td>-0.048</td>
<td>0.078</td>
<td>0.953</td>
<td></td>
<td></td>
<td>1.065</td>
</tr>
<tr>
<td></td>
<td>Model-2</td>
<td>0.130</td>
<td>0.086</td>
<td>-0.078</td>
<td>0.830</td>
<td>-0.001</td>
<td>0.091</td>
<td>0.859</td>
<td></td>
<td></td>
<td>1.114</td>
</tr>
<tr>
<td></td>
<td>Model-3</td>
<td>0.099</td>
<td>0.088</td>
<td>-0.203</td>
<td>0.654</td>
<td>-0.020</td>
<td>0.066</td>
<td>0.730</td>
<td></td>
<td></td>
<td>1.070</td>
</tr>
<tr>
<td></td>
<td>Model-4</td>
<td>0.625</td>
<td>0.147</td>
<td>1.676</td>
<td>2.624</td>
<td>0.231</td>
<td>0.324</td>
<td>1.057</td>
<td></td>
<td></td>
<td>1.045</td>
</tr>
<tr>
<td></td>
<td>Model-5</td>
<td>0.124</td>
<td>0.003</td>
<td>-0.843</td>
<td>0.859</td>
<td>-0.087</td>
<td>0.089</td>
<td>1.504</td>
<td></td>
<td></td>
<td>0.750</td>
</tr>
<tr>
<td></td>
<td>Model-6</td>
<td>0.100</td>
<td>0.150</td>
<td>-0.347</td>
<td>0.638</td>
<td>-0.037</td>
<td>0.066</td>
<td>0.784</td>
<td></td>
<td></td>
<td>1.183</td>
</tr>
<tr>
<td></td>
<td>Ensemble Mean</td>
<td>0.200</td>
<td>0.093</td>
<td>0.008</td>
<td>1.025</td>
<td>0.013</td>
<td>0.115</td>
<td>0.938</td>
<td></td>
<td></td>
<td>1.077</td>
</tr>
<tr>
<td>AOD simulations compared against VIIRS Deep Blue retrievals</td>
<td>Model-1</td>
<td>0.069</td>
<td>0.276</td>
<td>0.079</td>
<td>0.733</td>
<td>0.003</td>
<td>0.038</td>
<td>0.692</td>
<td></td>
<td></td>
<td>1.292</td>
</tr>
<tr>
<td></td>
<td>Model-2</td>
<td>0.099</td>
<td>0.314</td>
<td>0.820</td>
<td>1.255</td>
<td>0.053</td>
<td>0.074</td>
<td>0.755</td>
<td></td>
<td></td>
<td>1.280</td>
</tr>
<tr>
<td></td>
<td>Model-3</td>
<td>0.085</td>
<td>0.260</td>
<td>0.564</td>
<td>1.011</td>
<td>0.031</td>
<td>0.054</td>
<td>0.763</td>
<td></td>
<td></td>
<td>1.248</td>
</tr>
<tr>
<td></td>
<td>Model-4</td>
<td>0.592</td>
<td>0.285</td>
<td>4.042</td>
<td>4.614</td>
<td>0.291</td>
<td>0.317</td>
<td>0.912</td>
<td></td>
<td></td>
<td>1.186</td>
</tr>
<tr>
<td></td>
<td>Model-5</td>
<td>0.066</td>
<td>0.088</td>
<td>-0.666</td>
<td>0.752</td>
<td>-0.033</td>
<td>0.038</td>
<td>1.164</td>
<td></td>
<td></td>
<td>0.962</td>
</tr>
<tr>
<td></td>
<td>Model-6</td>
<td>0.060</td>
<td>0.319</td>
<td>0.223</td>
<td>0.694</td>
<td>0.012</td>
<td>0.036</td>
<td>0.620</td>
<td></td>
<td></td>
<td>1.350</td>
</tr>
<tr>
<td></td>
<td>Ensemble Mean</td>
<td>0.162</td>
<td>0.323</td>
<td>0.960</td>
<td>1.474</td>
<td>0.068</td>
<td>0.093</td>
<td>0.752</td>
<td></td>
<td></td>
<td>1.286</td>
</tr>
<tr>
<td>PM$_{2.5}$ simulations compared against AirNow observations</td>
<td>Model-1</td>
<td>8.600</td>
<td>0.378</td>
<td>0.250</td>
<td>0.621</td>
<td>2.151</td>
<td>5.318</td>
<td>0.538</td>
<td>15.317</td>
<td>67.053</td>
<td>2.187</td>
</tr>
<tr>
<td></td>
<td>Model-2</td>
<td>5.675</td>
<td>0.475</td>
<td>-0.110</td>
<td>0.450</td>
<td>-0.903</td>
<td>3.866</td>
<td>0.477</td>
<td>7.661</td>
<td>37.755</td>
<td>2.512</td>
</tr>
<tr>
<td></td>
<td>Model-3</td>
<td>13.543</td>
<td>0.287</td>
<td>0.470</td>
<td>0.891</td>
<td>3.938</td>
<td>7.686</td>
<td>0.663</td>
<td>12.969</td>
<td>86.406</td>
<td>1.762</td>
</tr>
<tr>
<td></td>
<td>Model-4</td>
<td>25.779</td>
<td>0.258</td>
<td>1.156</td>
<td>1.699</td>
<td>10.517</td>
<td>15.312</td>
<td>0.730</td>
<td>16.688</td>
<td>92.082</td>
<td>1.787</td>
</tr>
<tr>
<td></td>
<td>Model-6</td>
<td>7.261</td>
<td>0.267</td>
<td>-0.070</td>
<td>0.519</td>
<td>-0.795</td>
<td>4.485</td>
<td>0.542</td>
<td>11.076</td>
<td>49.938</td>
<td>2.264</td>
</tr>
<tr>
<td></td>
<td>Ensemble Mean</td>
<td>10.482</td>
<td>0.395</td>
<td>0.355</td>
<td>0.759</td>
<td>3.139</td>
<td>6.721</td>
<td>0.538</td>
<td>14.822</td>
<td>62.425</td>
<td>2.239</td>
</tr>
</tbody>
</table>

7.5 Ensemble Probability Forecast of PM$_{2.5}$ Exceedances

In this section, the ensemble exceedance probability forecasts (or binary prediction) were evaluated with the AirNow observed PM$_{2.5}$ exceedances. Overall, the ensemble exceedance forecast performed moderately in providing probability forecasts of dust-related PM$_{2.5}$ exceedances (concentration above 35 µg/m$^3$) during the 2021 Spring
Dust Season (January-March 2021). Figure 20 shows a contour map of the ensemble probability forecast values overlaid by the actual exceedances over the AirNow sites across the CONUS. The probability ranges from 20% (exceedances predicted by only one model) to 100% (exceedances predicted by all five models).

On March 16th, 2021, exceedances were observed at three AirNow monitoring sites in western Texas, in El Paso (56 and 55 µg/m³) and Socorro (54 µg/m³), and two sites in southern New Mexico, in Hobbs (40 µg/m³) and Santa Teresa (57 µg/m³) (marked as red filled circles). The ensemble exceedance probability was 20% (one model; light blue) and located in western Texas, partially over the downwind areas in eastern New Mexico, and along the southern Colorado-Kansas border. The ensemble exceedance probability reached 40% (two models; cyan) over relatively constrained areas in northwestern Texas to eastern New Mexico. The ensemble exceedance probability of 20% and 40% were predicted over the Southeast U.S. Coast, predominantly in Georgia. However, there were no exceedances observed by the AirNow sites in these areas. These false exceedances were consistent with the aforementioned overestimations of PM$_{2.5}$ by most participating models.

The ensemble exceedance forecast performance is shown as a time series plots of $aH$ and $aFAR$ in Figures 21a-21b. The average $aH$ and $aFAR$ over the study period are listed in Table 6. High $aH$ value and low $aFAR$ values suggest good agreement between model predictions and observations. The lowest ensemble probability of 20% shows the best $aH$ of 28.965 and the worst $aFAR$ of 94.936 on average, while the higher ensemble probability values show worse $aH$ and better $aFAR$. Since the low ensemble exceedance
probability forecast (20%) used only one member to predict exceedances, these simulated exceedances were not calibrated with other members. As a result, a few predicted exceedances were actually observed and the majority were false exceedances, leading the worst $aFAR$. In contrast, higher ensemble probabilities used more models to predict the exceedances performed more accurately and reliably in forecasting PM$_{2.5}$ exceedances on a smaller or local scale because their predicted exceedances have been calibrated with the co-existed exceedances predicted by the other models included in the ensemble. Therefore, the false exceedances were greatly reduced, and the predicted dust-related exceedances were constrained only in dust-active regions, resulting in better $aFAR$. In this case, the $aH$ of the high ensemble exceedance probability (80%-100%) forecasts equal to 0, referring to the failures of most members to predict the exceedances during dust storms. In addition, $aH$ of 0 may describe different scenarios depending on the $aFAR$. For instance, an ensemble probability of 100% showing average $aH$ and $aFAR$ values of 0 suggests no mutual exceedances predicted by all five individual models during the study period (Figure 21a), meaning no false alarms were counted (Figure 21b). On the other hand, the ensemble probability of 80% showed an average $aH$ value of 0 and average $aFAR$ value of 3.409, indicating that there were mutual exceedances predicted by four individual models that were false alarms as none of them were observed by AirNow sites. All results indicate that the ensemble underperformed in providing high exceedance probability (80-100%) forecasts over areas affected by high concentrations of dust PM$_{2.5}$ above the NAAQS health standard.
In practice, the accuracy of the exceedance probability forecast depends on the original spatial resolutions of each ensemble member. The exceedances simulated by the global models generally cover larger areas compared to the regional models, even after being interpolated to a higher spatial resolution. Using the multi-model ensemble approach generally reduces the discrepancies between the spatial resolutions of the ensemble members. Although the ensemble was able to generate only low-medium (20%-60%) exceedance probability forecasts during dust storms, it can probably be used to provide hazardous areas during dust storms in addition to the ground observations, as shown in Figure 20.

The major challenge in dust associated PM$_{2.5}$ exceedance forecasting is the occurrence of dust storms in active dust regions, which are relatively small areas compared to other air pollution sources and are limited in duration, generally ranging from a few minutes to several hours. Therefore, dust-related exceedances in the active dust regions are frequently underpredicted and the remaining predicted exceedances elsewhere could be false.
Figure 20. Ensemble probability forecast of PM$_{2.5}$ exceedances on March 16th, 2021 (during the 2021 Spring Dust Season). Foreground colors indicate the probability values ranging from 20% (one out of five models forecasts the PM$_{2.5}$ exceedance) (light blue) to 100% (all five models forecast the PM$_{2.5}$ exceedances) (red). The PM$_{2.5}$ exceedances observed by the AirNow sites are displayed in the red/green circles (red means an exceedance recorded by the monitor, and green means no exceedance recorded).
Figure 21. Time series plots of \( aH \) (a) and \( aFAR \) (b) values during the 2021 Spring Dust Season (January-March 2021) for the ensemble probability of PM\(_{2.5}\) exceedance forecast. Ensemble probability values range from 20% (one out of five models) to 100% (all five models).

Table 6. Averaged \( aH \) and \( aFAR \) values of ensemble probability of PM\(_{2.5}\) exceedance forecast during the 2021 Spring Dust Season (January-March 2021), comparing between simulated PM\(_{2.5}\) exceedances and observed PM\(_{2.5}\) exceedances obtained from the AirNow ground monitoring network.

<table>
<thead>
<tr>
<th>Ensemble Probability</th>
<th>( aH )</th>
<th>( aFAR )</th>
</tr>
</thead>
<tbody>
<tr>
<td>20%</td>
<td>28.965</td>
<td>94.936</td>
</tr>
<tr>
<td>40%</td>
<td>17.712</td>
<td>60.029</td>
</tr>
<tr>
<td>60%</td>
<td>4.659</td>
<td>16.032</td>
</tr>
<tr>
<td>80%</td>
<td>0.000</td>
<td>3.409</td>
</tr>
<tr>
<td>100%</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

\( aH \): area hit rate; \( aFAR \): area false alarm ratio
CHAPTER 8: CONCLUSIONS

8.1 Conclusion and Recommendation for Future Work

Wildfires and dust storms are important emission sources that contribute a large amount of aerosols to the atmosphere, leading to hazardous air quality, which 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 the North America ensemble wildfire and dust air pollution forecasts of AOD and PM$_{2.5}$ in order to predict wildfire and dust storm effects on AOD and surface PM$_{2.5}$. The multi-model ensemble forecasts were built using three (for wildfire) and two (for dust storm) regional models, one global ensemble model, and three global models operated by NASA, NOAA, NRL, and George Mason University (GMU). These models include the GMU-CMAQ, NACC-CMAQ, HYSPLIT, ICAP-MME, GEFS-Aerosols, GEOS-5, and NAAPS. Our ensemble forecast reproduces daily forecasts of AOD and PM$_{2.5}$ as well as the ensemble probability forecast for wildfire and dust related PM$_{2.5}$ exceedances (24-hr average concentrations $>$35 μg/m$^3$) on a horizontal grid resolution of 12 km×12 km over the CONUS during the 2020 Gigafire events (August-September 2020) in the western U.S. and during the Spring Dust Season (January-March 2021) in the Chihuahuan Desert.
The performance of the ensemble forecasting for AOD and PM$_{2.5}$ during wildfire and dust storm events was evaluated with VIIRS enhanced Dark Target (DT) over dark and bright surfaces, VIIRS Deep Blue (DB), and MAIAC AOD products, and AirNow surface PM$_{2.5}$ observations by calculating a suite of statistical metrics ($RMSE$, $CORR$, $MB$, $ME$, $NMB$, $NME$, and $FB$) and an overall rating ($RANK$). In addition, two discrete categorical metrics (area hit rate; $aH$ and area false alarm ratio; $aFAR$) were employed to measure the performance of ensemble mean and ensemble probability in predicting the exceedances of the National Ambient Air Quality Standards (NAAQS) for PM$_{2.5}$ (concentrations $>35$ μg/m$^3$) during the wildfires and dust storms.

For the wildfire case, overall, the statistics results suggested the ensemble mean significantly reduces the biases and uncertainties in the wildfire air pollution forecast and produces more persistently reliable forecasts during the study period compared to the individual forecasts. For AOD forecasts, the ensemble mean was able to improve model performance, as indicated by the mean bias values greatly reduced to $-0.08$ and $-0.04$ based on the comparisons with VIIRS DT and MAIAC, the strongest correlations at $0.57$ from a range of $0.30$-$0.53$ (VIIRS DT) and $0.62$ from a range of $0.35$-$0.56$ (MAIAC). The ensemble mean also achieved the best ($1.45$ from a range of $0.95$-$1.44$) and second best overall $RANK$ among seven members compared to VIIRS DT and MAIAC. For the forecasts of surface PM$_{2.5}$, the ensemble mean outperformed all individual models, with the mean bias reduced to $7.40$, strongest correlation at $0.60$ from a range of $0.43$-$0.54$, the lowest fractional bias $0.54$ from a range of $0.55$-$1.32$, the highest area hit rate at $87\%$ from a range of $40\%$-$82\%$, and the best overall $RANK$ ($2.83$ from a range of $2.40$-$2.81$).
among six members. 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 area false alarm ratio 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 warning alerts when the PM$_{2.5}$ concentrations exceed the NAAQS health standard level (concentrations >35 μg/m$^3$) during wildfires. Although the evaluation result suggested that the ensemble is capable of reducing bias and uncertainties in the model forecasts, predicted AOD and surface PM$_{2.5}$ are frequently subject to be decoupled due to the vertical distribution of the smoke particles. Nevertheless, the relatively high negative biases and positive biases values of the ensemble forecast in the middle of September 2020 demonstrate the underestimations of AOD and overestimations of PM$_{2.5}$ during intense wildfires, which may have been influenced by the unusually thick smoke that in turn caused large errors in emission estimation and plume injection height calculation (Carter et al., 2020; Pan et al. 2020b; and Ye et al. 2021). Furthermore, the variety of model simulations as well as meteorology and emissions inputs (both initial and boundary conditions) data sets can take into account the uncertainty in the ensemble forecasting.

For the dust storm case, overall, the statistical results suggested that the ensemble mean shows the ability to moderately reduce biases in the AOD and PM$_{2.5}$ predictions over the active dust and downwind areas and provide fairly reliable forecasts during the dust events. For AOD forecasts, the ensemble mean improved model forecasts less
successfully than expected, as demonstrated by the mean bias being minimized to 0.01 and 0.07 (based on the comparisons with VIIRS DT and VIIRS DB, respectively), the strongest correlation at only 0.32 (based on VIIRS DB), and having the third highest $RANK$ among six members (for both VIIRS DT and DB). For surface PM$_{2.5}$, the ensemble forecasts underperformed, as indicated by mean bias slightly decreasing to 3.32, a fairly weak correlation (0.40), low area hit rate (14.82%), and the third highest $RANK$ among five members. In terms of dust associated PM$_{2.5}$ exceedance probability forecasts (binary prediction), the ensemble unsuccessfully generated high probability forecasts of PM$_{2.5}$ exceedance during dust storm events. Instead, it frequently predicted exceedances with low-medium probabilities (20-60%) during the dust events. However, the ensemble probability still has the useful capability to estimate the hazardous areas affected by the dust PM$_{2.5}$ exceedances, especially over areas which are generally difficult to establish air quality monitoring sites.

The predominant obstacles for the dust storm ensemble forecasts are the excessive overestimations and underestimations from the ensemble members, primarily in the middle of March when extreme dust storms occurred in the Chihuahuan Desert. These biases arise from the model simulation errors caused by large uncertainties in dust emissions and other emissions like prescribed fire and anthropogenic emissions, dust transport and deposition processes, different applications of dust parameterizations in dust scheme algorithms, and meteorological fields adopted differently in each individual model. In addition, the natural behavior of the local dust events in the western U.S. causes issues in dust forecasting because they
are generally limited in duration, small-scale wind circulations, and restricted to a particular region.

In essence, using the ensemble approach can reduce biases in air pollution forecasts and reasonably improve the model predictability during extreme events such as wildfires and dust storms. However, the reliability of the ensemble forecast is still subject to types of extreme events due to different emission sources as well as initial and boundary meteorological conditions.

The development and evaluation of the multi-model ensemble wildfire and dust storm air pollution forecast for the 2020 Gigafire events and the 2021 Spring Dust Season presented here are still at the early stage of deploying the North America ensemble wildfire and dust storm forecast. Comparisons between the ensemble and the individual models represented in this study will be used to investigate differences between models as an attempt to identify the uncertainties in emission and meteorology inputs, as well as in chemical transport/dispersion model simulations. Findings from this pilot study will be used to improve forecasting performance of the ensemble mean and each individual model. Our next step is to extend the multi-model ensemble forecast approach to other periods, including the 2021 fire and dust storm seasons. Finally, the qualified ensemble forecast will be used to improve real-time wildfire forecasting systems over North America to support key decision-making on air quality at local, national, and international levels.
APPENDIX: THE CONFIGURATION OF PARTICIPATING MODEL

Table A1. The configuration of participating models included in the ensemble forecasting.

<table>
<thead>
<tr>
<th>Air Quality/Dispersion Model</th>
<th>Operational/Research Center</th>
<th>Forecast Products</th>
<th>Domain(s)</th>
<th>Meteorology Data</th>
<th>Grid Spacing</th>
<th>Initial time</th>
<th>Output Frequency</th>
<th>Fire Detection Information</th>
<th>Emission product/Algorithm</th>
<th>Dust Algorithm</th>
<th>Plume Rise Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMU-CMAQ (CMAQ v5.3.1)</td>
<td>Air Quality Group at George Mason University</td>
<td>PM$_{2.5}$, O$_3$, NO$_x$, AOD, etc.</td>
<td>CONUS and its surrounding areas</td>
<td>GFS products applied to WRF v4.2</td>
<td>12kmx12km</td>
<td>18 UTC previous day</td>
<td>Hourly</td>
<td>GHGEPx from VIIRS and MODIS</td>
<td>BEIS, MOVES, 2016v1, Grong (2003) sea-spray emission, applied to SM03K1=4.7</td>
<td>FENOCHA dust scheme</td>
<td>Sofiev (2012)</td>
</tr>
<tr>
<td>NACC-CMAQ (CMAQ v5.3.1)</td>
<td>NOAA Air Resources Laboratory (ARL)/GMU</td>
<td>PM$_{2.5}$, O$_3$, NO$_x$, AOD, etc.</td>
<td>CONUS and its surrounding areas</td>
<td>GFSv16 with FV3</td>
<td>12kmx12km</td>
<td>12 UTC</td>
<td>Hourly</td>
<td>GHGEPx from VIIRS and MODIS</td>
<td>BEISv3.6.1-BELDS, MOVES, NEED016v1</td>
<td>FENOCHA dust scheme, SOILGRIDS 2017 soil factors, Prickett et al (2012) surface roughness</td>
<td>Briggs</td>
</tr>
<tr>
<td>HYSEPL (HYSEPL v5.1.6)</td>
<td>NOAA Air Resources Laboratory (ARL)</td>
<td>PM$_{2.5}$, AOD</td>
<td>CONUS and its surrounding areas</td>
<td>WRF-ARW</td>
<td>0.15°x0.15°</td>
<td>00 UTC</td>
<td>Hourly</td>
<td>NOAA Hazard Mapping System (HMS)</td>
<td>USFS BlueSky</td>
<td>HYSEPL threshold friction velocity dust scheme</td>
<td>Briggs</td>
</tr>
<tr>
<td>GEOS-5 (GEOS-5 v5.27.1)</td>
<td>NASA Goddard Space Flight Center (GSC)</td>
<td>PM$_{2.5}$, AOD</td>
<td>Global</td>
<td>Near-real time assimilation (DAAS), 10-days forecast at 00z, and 5-days forecast at 12z</td>
<td>12kmx12k, with output at 0.25°x0.3125°</td>
<td>00 UTC</td>
<td>Hourly</td>
<td>NOAA Hazard Mapping System (HMS)</td>
<td>Prognostic emissions of dust and aerosol, smoke - QFD, HTAP, EDGAR, MEGAN</td>
<td>GEOS/GOCART</td>
<td>Briggs</td>
</tr>
<tr>
<td>GEFS-Aerosols</td>
<td>NOAA ARL/NOAA CSL/NOAA GSL/GMU</td>
<td>PM$<em>{2.5}$, PM$</em>{10}$, PM$_{10}$, AOD, OC, BC, Dust, Sea Salt, SO$_4$</td>
<td>Global</td>
<td>FV3GFS</td>
<td>0.25°x0.25°</td>
<td>00 UTC</td>
<td>3 Hourly</td>
<td>MODIS</td>
<td>CEDS,HTAP</td>
<td>FENOCHA dust scheme, SOILGRID 2017 soil factors, Prickett et al (2012) surface roughness</td>
<td>Briggs</td>
</tr>
<tr>
<td>ICAP-MMIE</td>
<td>Naval Research Laboratory (NRL)</td>
<td>AOD</td>
<td>Global</td>
<td>Varies</td>
<td>1°x1°</td>
<td>00 UTC</td>
<td>6 hourly</td>
<td>Varies</td>
<td>Varies</td>
<td>Varies</td>
<td>1D cloud</td>
</tr>
<tr>
<td>NAAPS (nap0440)</td>
<td>Naval Research Laboratory (NRL)</td>
<td>PM$_{2.5}$, AOD, visibility</td>
<td>Global</td>
<td>Global Meteorological fields from NAVGEM</td>
<td>0.333°x0.333°</td>
<td>00 UTC</td>
<td>3 hourly</td>
<td>MODIS</td>
<td>Dust, sea salt, anthropogenic and biogenic fine mode (ABF)</td>
<td>Westphal et al. 2009</td>
<td>Varies</td>
</tr>
</tbody>
</table>
REFERENCES

http://dx.doi.org/10.1088/1748-9326/aabf20


gridded soil information based on machine learning. *PLOS ONE, 12*(2), e0169748. https://doi.org/10.1371/journal.pone.0169748


part of the “Deep Blue” aerosol project. *Journal of Geophysical Research Atmospheres, 123*(1), 380–400. [https://doi.org/10.1002/2017jd027412](https://doi.org/10.1002/2017jd027412)


Textor, C., Schulz, M., Guibert, S., Kinne, S., Balkanski, Y., Bauer, S., Berntsen, T., Berglen, T., Boucher, O., Chin, M., Dentener, F., Diehl, T., Feichter, J., Fillmore, D.,


95


130(12), 2905–2916.  
https://doi.org/10.1175/15200493(2002)130<2905:TDVAWS>2.0.CO;2

https://doi.org/10.1002/2016MS000723

https://doi.org/10.1029/2009GL038823

https://doi.org/10.1002/qj.3497

https://doi.org/10.5194/acp-21-14427-2021

https://doi.org/10.1117/1.JRS.12.042609

https://doi.org/10.5194/gmd-2021-378

https://doi.org/10.1029/2012JD017459

97


BIOGRAPHY

Peewara Makkaroon is a Master’s student in Earth Systems Science, a shared program between the Department of Atmospheric, Oceanic, and Earth Sciences and the Department of Geography and Geoinformation Science at George Mason University (GMU). In 2018, she received her Bachelor of Science from Chulalongkorn University, Bangkok, Thailand. Until Fall 2020, she worked as a geophysicist at PinPoint Geophysics, Bangkok, Thailand. She started working as a Graduate Research Assistant at GMU in Summer 2021 under the supervision of Dr. Daniel Tong. Her primary research area is the air quality modeling associated with dust and wildfire events. In addition, she has general expertise in Geology, Atmospheric Science, and Geographic Information System.