Model Evaluation

Yang Zhang, Khanh Do, and Daniel Schuch
Northeastern University

September 25, 2024

World Meteorological Organization's

"Training course on Seamless Prediction of Air Pollution in Africa"

Online webinars, 12-25 September 2024

Outline

- **Importance**
- **Model Evaluation**
  - Datasets for Model Evaluation
  - Types of Model Evaluation
  - Current Status of Model Evaluation
- **Preliminary Evaluation of Initial WRF-Chem Application over Africa**
  - Specific Datasets Used for Model Evaluation
  - Evaluation Metrics and Protocols
  - Meteorological Evaluation (WRF only, Jan-April 2023)
  - Chemical Evaluation (WRF-Chem, Jan, 2023)
- **Summary**

Major sources: Zhang et al., 2006a,b, 2019, Zhang (2024), Dennis et al., 2010; Emery and Tai, 2001; Emery et al., 2017

Why is Model Evaluation Important?

- Evaluate model performance skill in terms of accuracy and reliability
- Assess if the ambient air quality meets the air quality standards
- Identify model biases and missing processes for potential model improvement
- Perform accurate source apportionment to support decision-making
- Evaluate model sensitivity to model parameters and processes
- Evaluate uncertainties in model inputs, representations, and configurations
- Establish creditable baseline for projection of future air quality
- Deepen process-level understanding of sciences

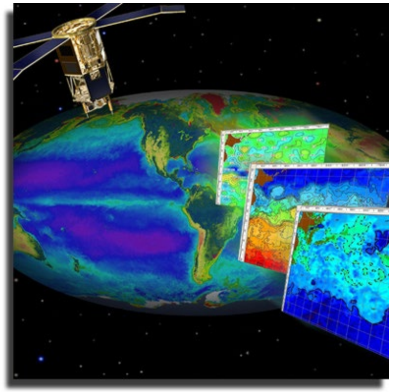
Datasets for Model Evaluation (Zhang, 2024)

- Emission Measurements
- Deposition Measurements
- Ground- and Upper Air Meteorological and Chemical Concentration Observations
- Satellite-Based Observations
- Reanalysis Datasets
- Relevant data generated using data fusion and ML

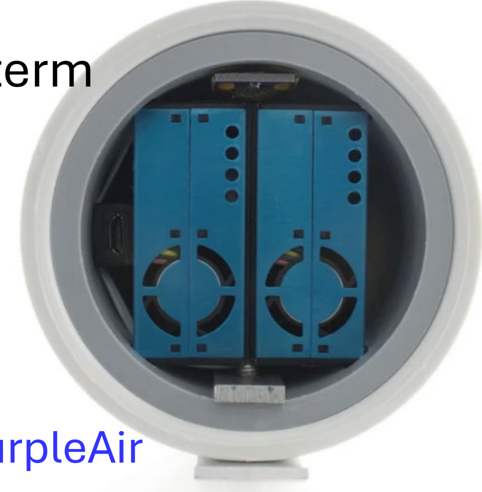
Merits and Limitations of Datasets

- **Ground truth from ground monitoring stations**
  - Most accurate
  - Sparse and limited access
- **Low-cost air quality sensors**
  - PurpleAir, Clarity, MODULAIR Air
  - Low-cost sensor evaluation: https://www.aqmd.gov/aq-spec
  - Require collocation and calibration
- **Satellite products (e.g., gaseous column abundance, AOD)**
  - Moderately accurate
  - Widely available, but requires extra steps for evaluation
- **Re-analysis data**
  - Coarse resolution but long-term
  - Acceptable quality
  - Widely available
- **ML/data fusion-based data**
  - high resolution, high fidelity
  - Limited time period

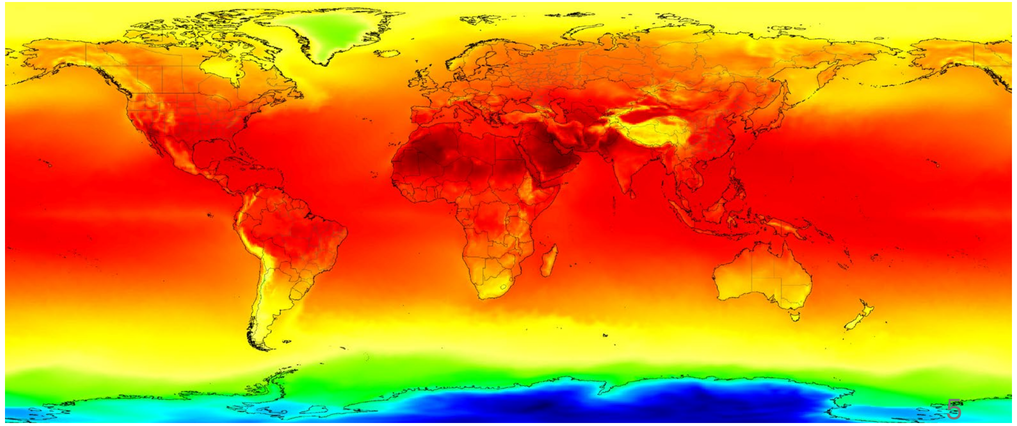
https://www.aqmd.gov



https://earth.gsfc.nasa.gov/climate/data/deep-blue



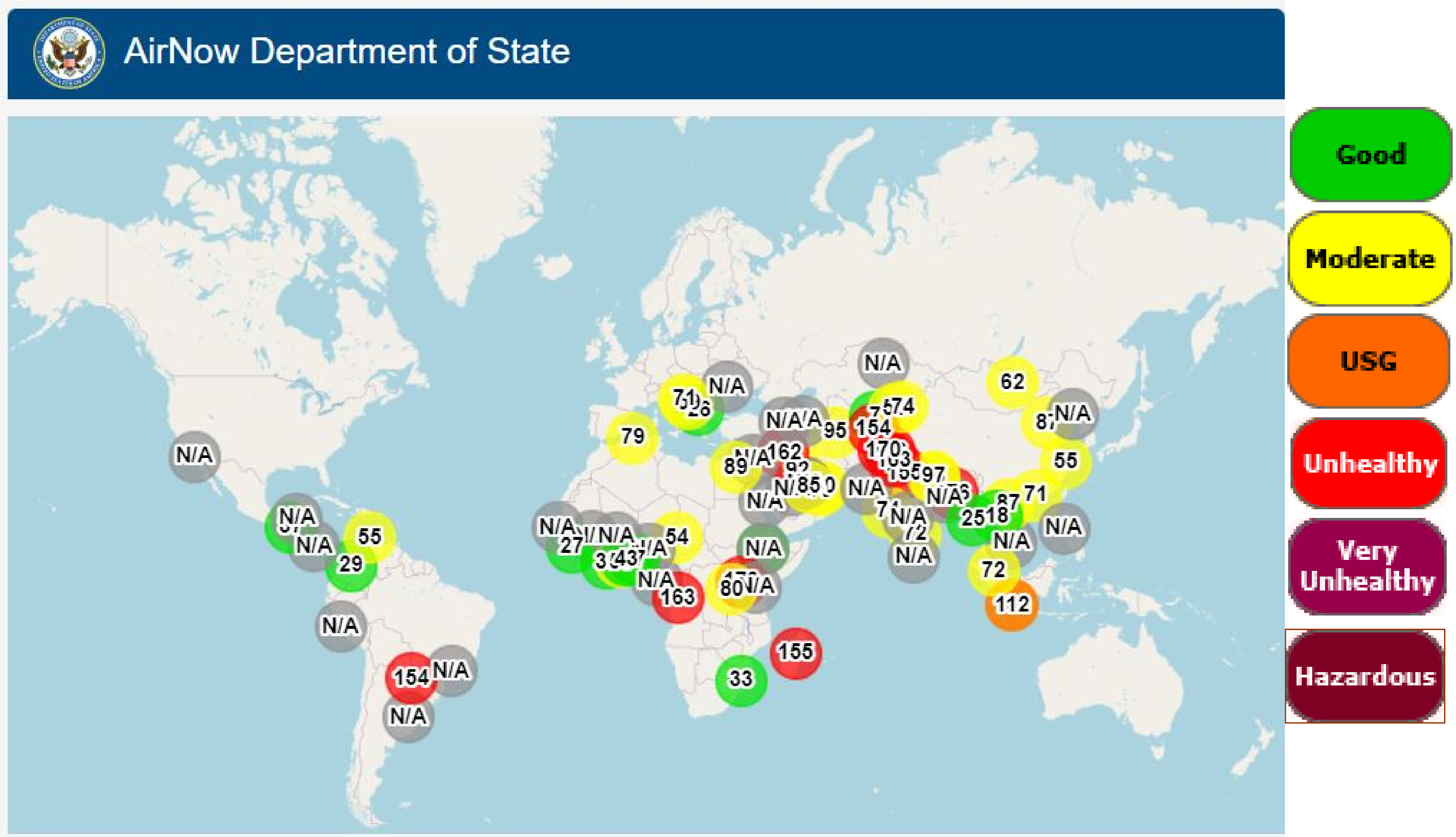
PurpleAir



https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/datacube/

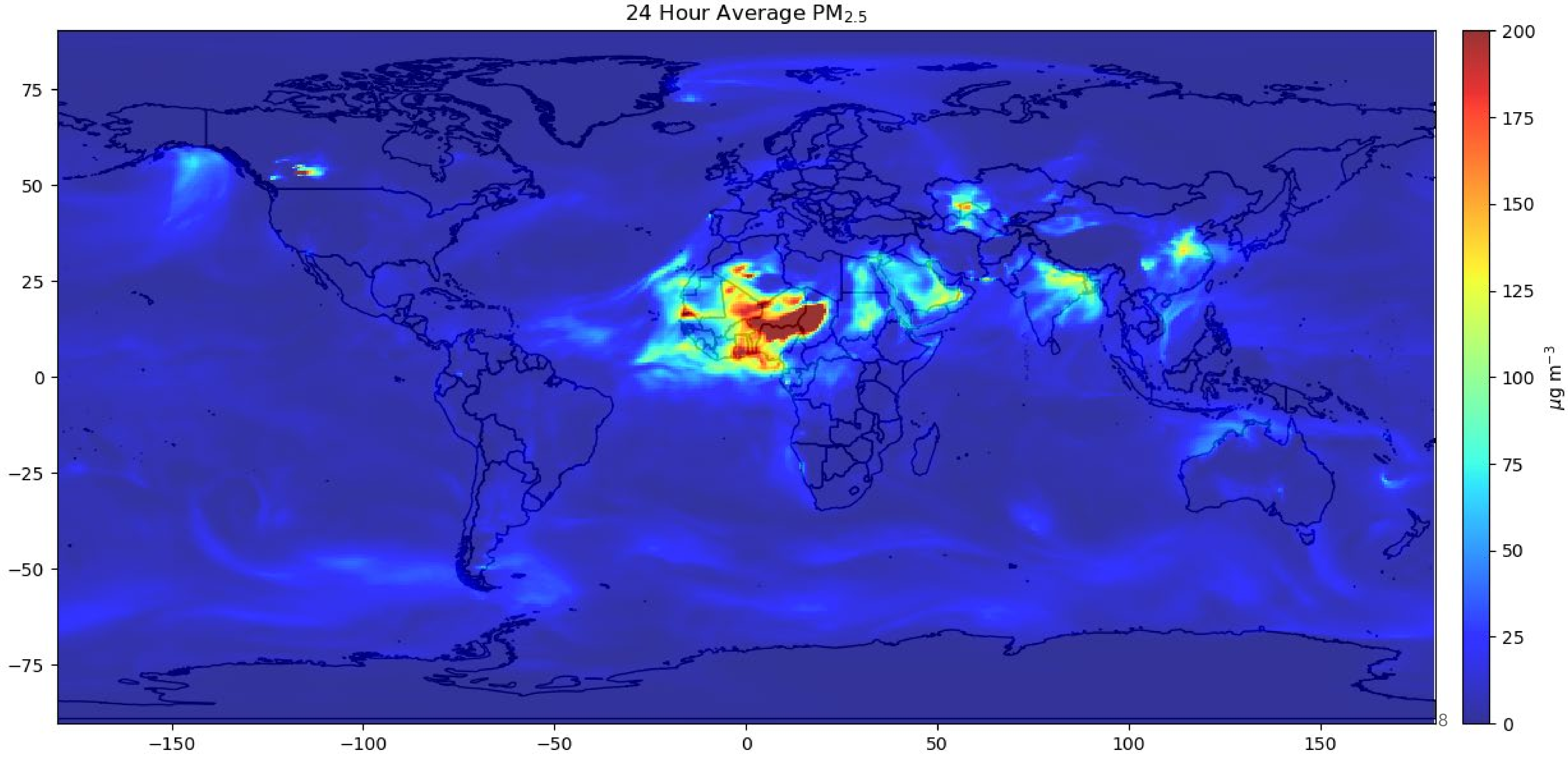
AirNow International

(https://www.airnow.gov/international/us-embassies-and-consulates/)



AERONET (https://aeronet.gsfc.nasa.gov/)

$PM_{2.5}$ from MERRA-2
(https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/)



Types of Model Performance Evaluation

Operational: Assessing the main output variable (e.g., T, WS, O_3, and $PM_{2.5}$)

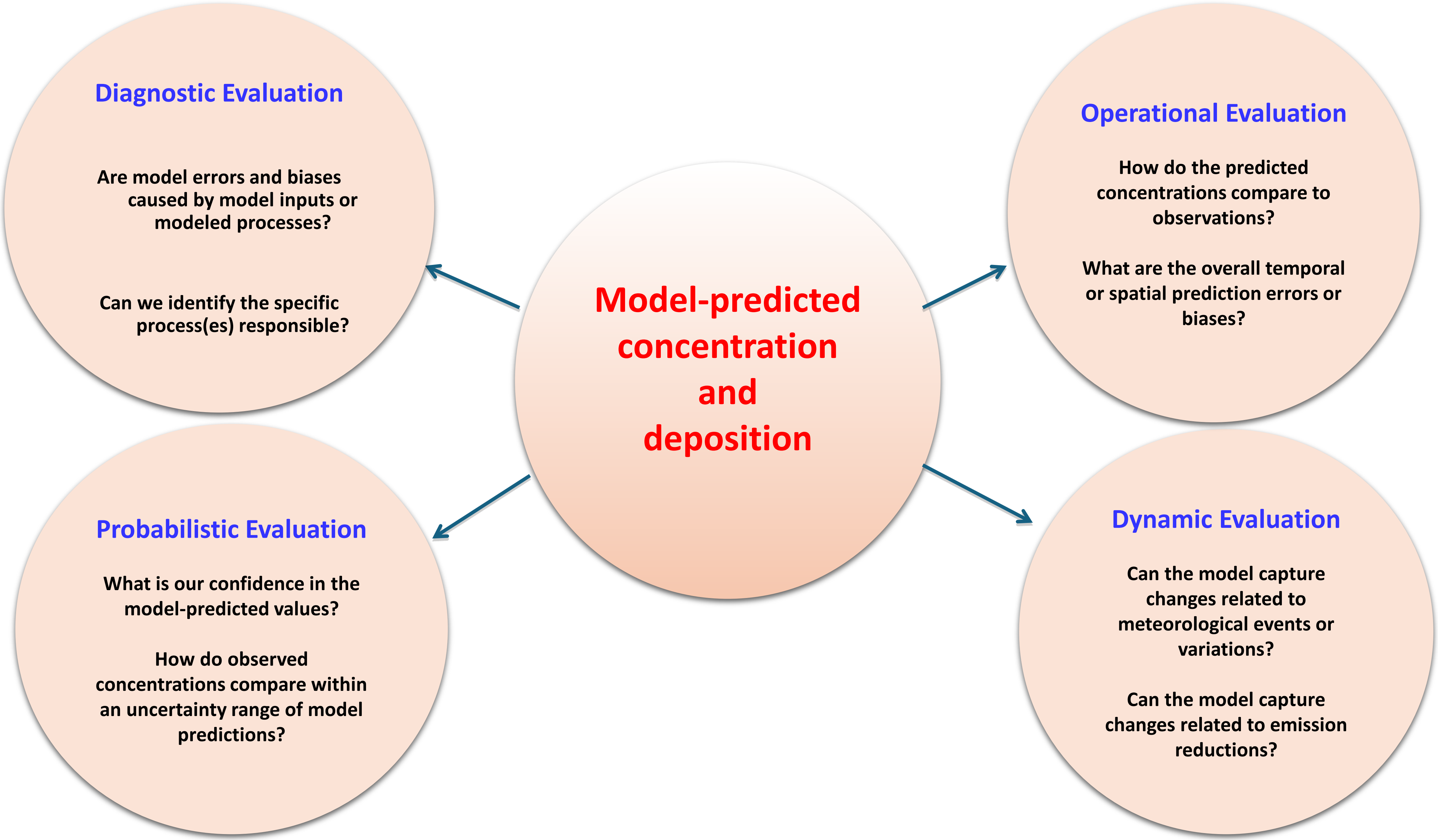
Discrete and Categorical (CW-AQF)

Diagnostic: Assessing the precursors/oxidants or components for the main output variable, model inputs, and major processes and parameters

Mechanistic: Assessing the responses of the output variables to changes in input variables and model parameters

Probabilistic: Assessing uncertainties in model outputs and observations

Model Evaluation Framework and Objectives (Dennis et al., 2010)



Schematic Representation of the Four Levels of Model Evaluation (Zhang, 2024)

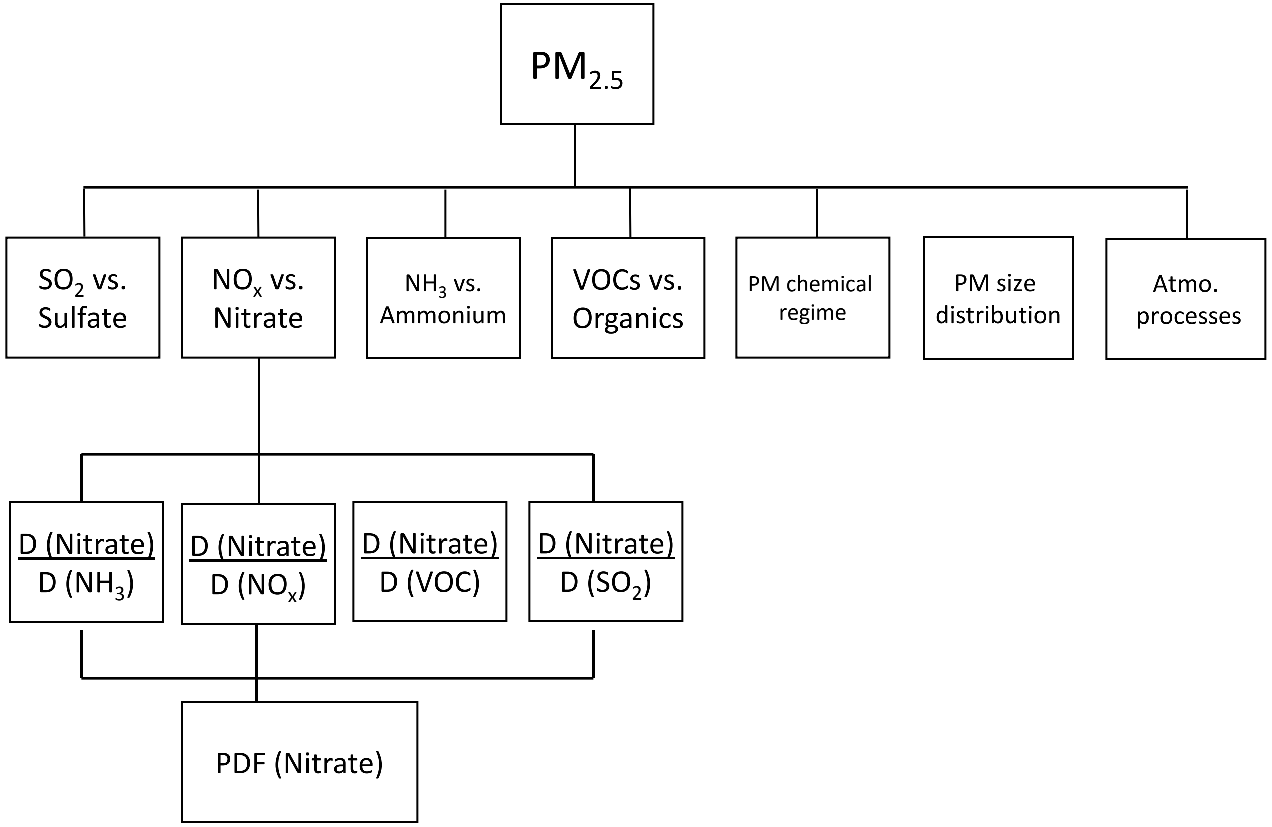
Model Performance Evaluation | **Model Analysis Technique**

Operational — Output Analysis

Diagnostic — Process Analysis

Mechanistic — Sensitivity Analysis

Probabilistic — Uncertainty Analysis



PDF – probability distribution function

Operational Discrete Evaluation (Zhang, 2024)

- **Variables evaluated**
  - **All raw outputs: meteorological variables, concentrations of gases and $PM_{2.5}$**
  - **Processed variables: column abundance, species ratios**
- **Statistics commonly used**
  - **Accuracy of peak (matched and unmatched in space)**
  - **Bias (or fractional bias)**
  - **Gross error (or fractional gross error)**

The mean bias (MB)

$$MB = \frac{1}{N}\sum_{i=1}^{N}(M_i - O_i) = \overline{M} - \overline{O} \qquad \overline{M} = (1/N)\sum_{i=1}^{N} M_i \qquad \overline{O} = (1/N)\sum_{i=1}^{N} O_i$$

The mean error (MAGE)

$$MAGE = \frac{1}{N}\sum_{i=1}^{N}|M_i - O_i|$$

The mean normalized bias (MNB)

$$MNB = \frac{1}{N}\sum_{i=1}^{N}[(M_i - O_i)/O_i] = \frac{1}{N}\sum_{i=1}^{N}(M_i/O_i - 1)$$

The mean normalized gross error (MNE)

$$MNE = \frac{1}{N}\sum_{i=1}^{N}[(|M_i - O_i|)/O_i]$$

The normalized mean bias (NMB)

$$NMB = [\sum_{i=1}^{N}(M_i - O_i)]/\sum_{i=1}^{N} O_i = (\frac{\overline{M}}{\overline{O}} - 1)$$

The normalized mean gross error (NME)

$$NME = [\sum_{i=1}^{N}|M_i - O_i|]/\sum_{i=1}^{N} O_i = MAGE/\overline{O}$$

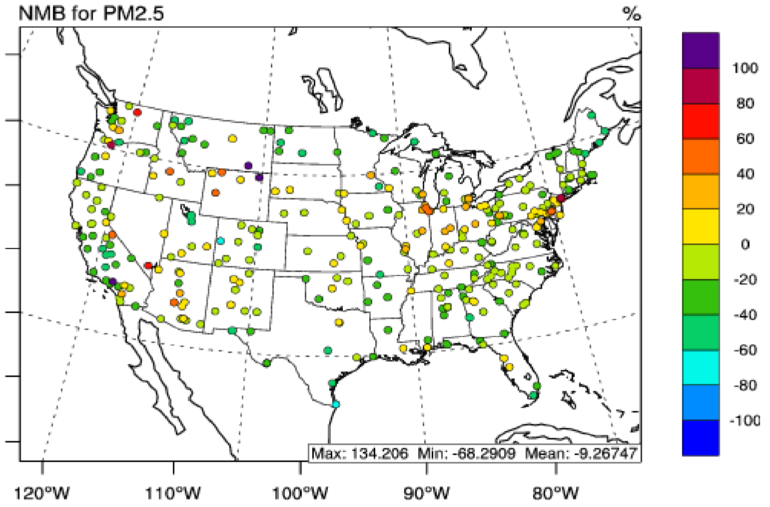
where N is the number of samples (by time and/or location), M_i and O_i are values of model prediction and observation at time and location , respectively.

Operational Evaluation of $PM_{2.5}$ (2006)

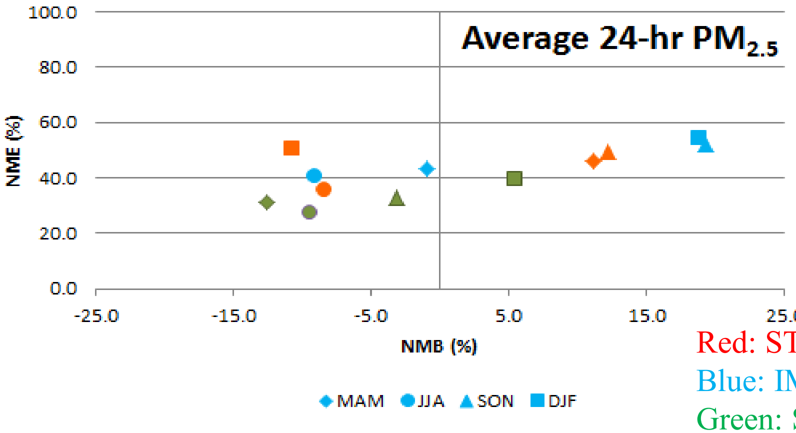
(IMPROVE, STN, SEARCH) (Yahya et al., 2014)

JJA NMB Spatial Distribution

24-h Avg. $PM_{2.5}$

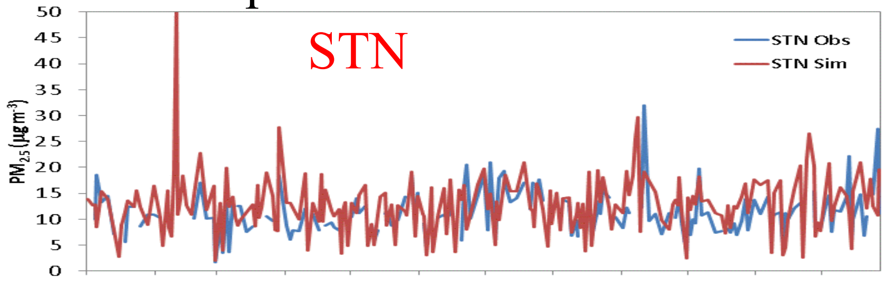


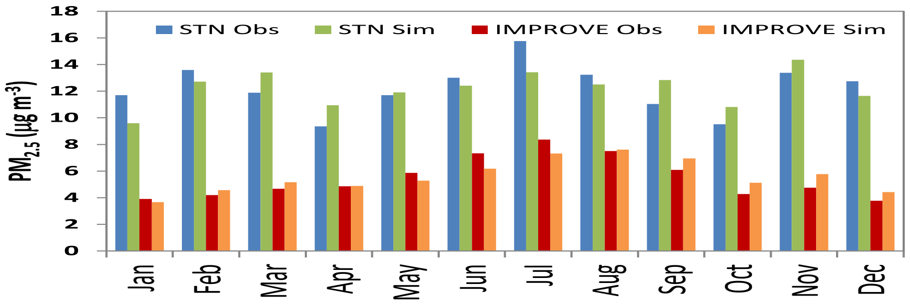
Soccer plots of NMB vs. NME



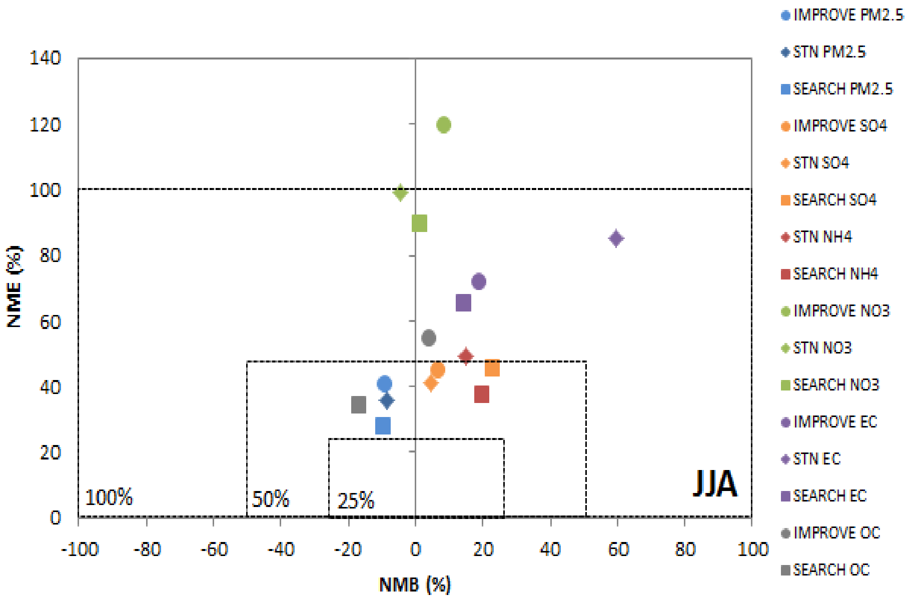
Red: STN
Blue: IMPROVE
Green: SEARCH

Temporal Variations



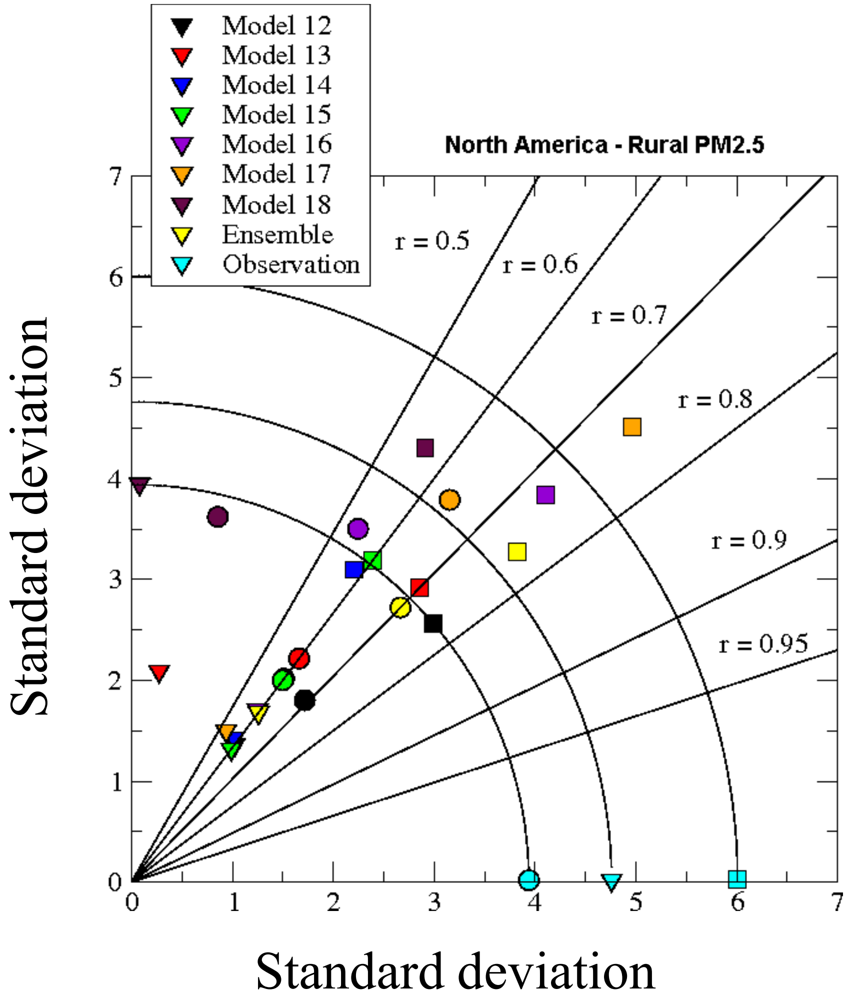


Soccer plots of $PM_{2.5}$ and species



Taylor Diagrams for $PM_{2.5}$ Performance at Rural Sites

(Solazzo et al., 2012)



- Each symbol indicates a different model run
- The position of the symbol on the diagram indicates:
  - the correlation between observations and model (as angle counter-clockwise from the "east" position"
  - the ratio of modeled-to-observed standard deviation (radial distance from the origin)
  - the centered pattern RMSE (distance from light blue symbols on the horizontal axis

(triangle: domain1; circle: domain2; square: domain3)

Operational Categorical Evaluation for CW-AQF Model (Zhang, 2024)

- **Accuracy (A)**
  Percentage of forecasts that correctly predict an exceedance or a nonexceedance
- **Critical Success Index (CSI)**
  Indicate how well actual exceedances are predicted, accounting for both missed events and false alarms
- **Probability Of Detection (POD)**
  Percentage of actual exceedances that are forecasted, accounting for only missed events
- **Bias (B)**
  Judges if forecasts are underpredicted (< 1) or overpredicted (> 1)
- **False Alarm Ratio (FAR)**
  Measures the percentage of times an exceedance was forecasted when none occurred

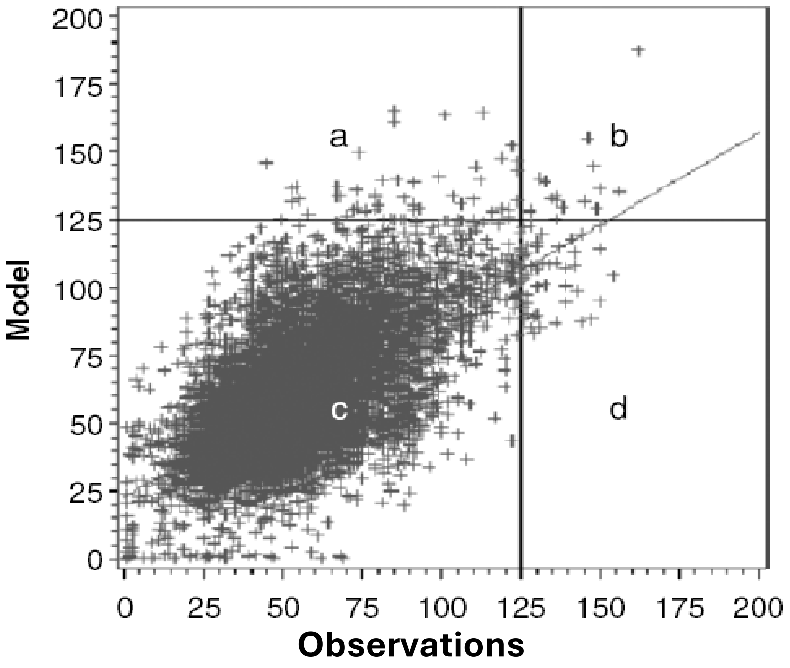
$$A = \left(\frac{b+c}{a+b+c+d}\right) \times 100\%$$

$$\mathrm{CSI} = \left(\frac{b}{a+b+d}\right) \times 100\%$$

$$\mathrm{POD} = \left(\frac{b}{b+d}\right) \times 100\%$$

$$B = \left(\frac{a+b}{b+d}\right)$$

$$\mathrm{FAR} = \left(\frac{a}{a+b}\right) \times 100\%$$

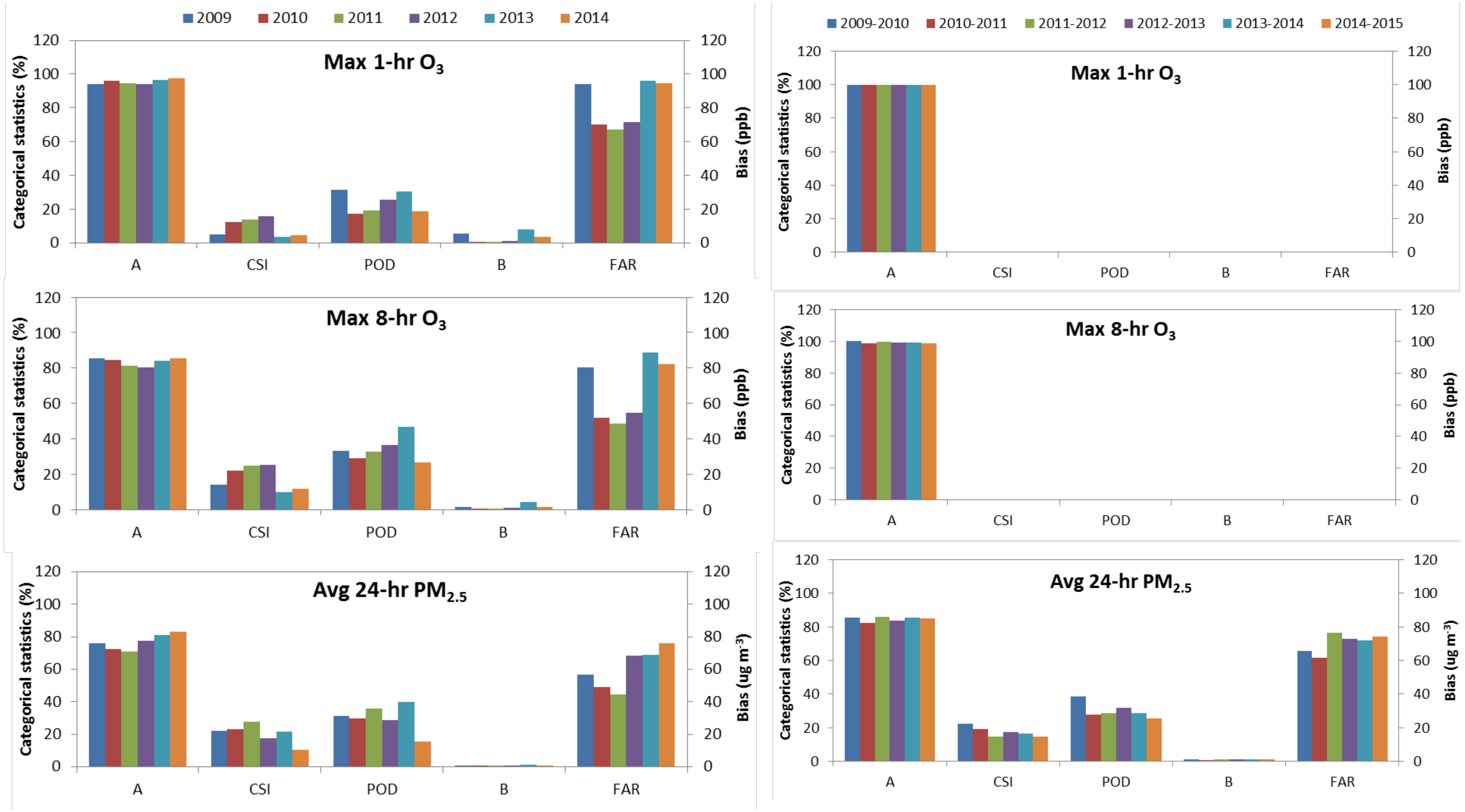


(Figure from Kang et al., 2005)

Categorical Evaluation Against AIRNow (2009-2014) (Zhang et al., 2016)

O_3 Season

Winter Season



- Overall good performance in terms of A and B, but with relatively low CSI and POD and high FAR

Diagnostic Evaluation (Zhang, 2024)

- **Analyses of PM chemical composition**

  Sulfate, nitrate, ammonium, elemental carbon (EC), organic carbon (OC), total nitrate (HNO_3+ PM nitrate), and total ammonium (NH_3 +PM ammonium)
- **Analyses of precursors of secondary PM**

  Primary precursors (SO_2, NO, NO_2, HNO_3, NH_3, and VOC)
  and oxidants and radicals (O_3, OH, NO_3, and H_2O_2)
- **Analyses of shorter time average concentrations**

  Nitrate, ammonium, and OC for diurnal variation; seasonal variation for annual PM
- **Analyses of light extinction**

  Scattering and absorption
- **Analyses of mass fluxes and governing processes**

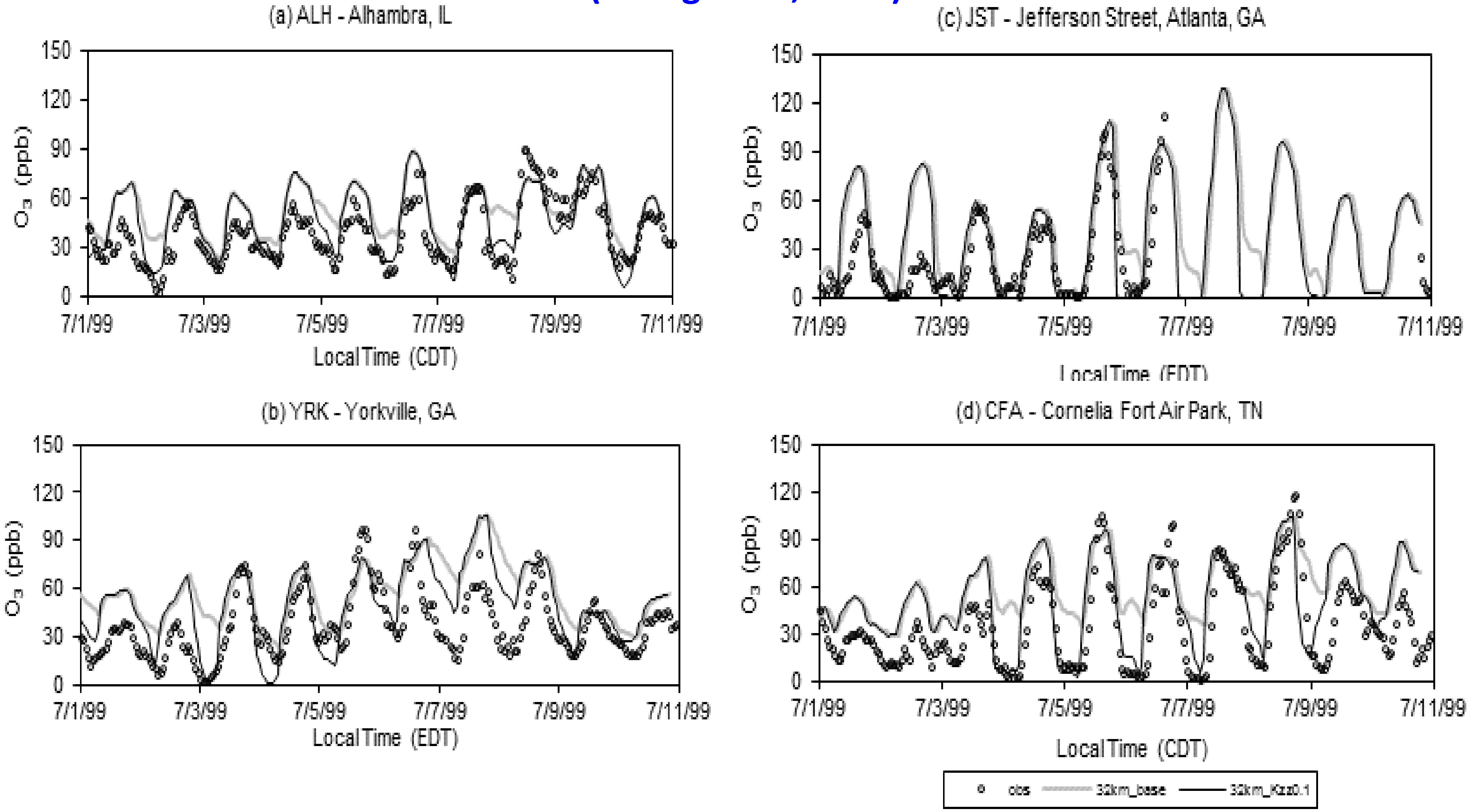
  Emissions, transport, transformation, and dry and wet deposition fluxes
- **Analyses of model inputs and parameters**

  Boundary conditions, rate coeff., vertical eddy diffusivity
- **Analyses of PM size distribution**

  modes (peaks and standard deviations), size intervals, and distribution shapes

Impact of the floor value of K_{zz} on O_3 during the SOS99 episode

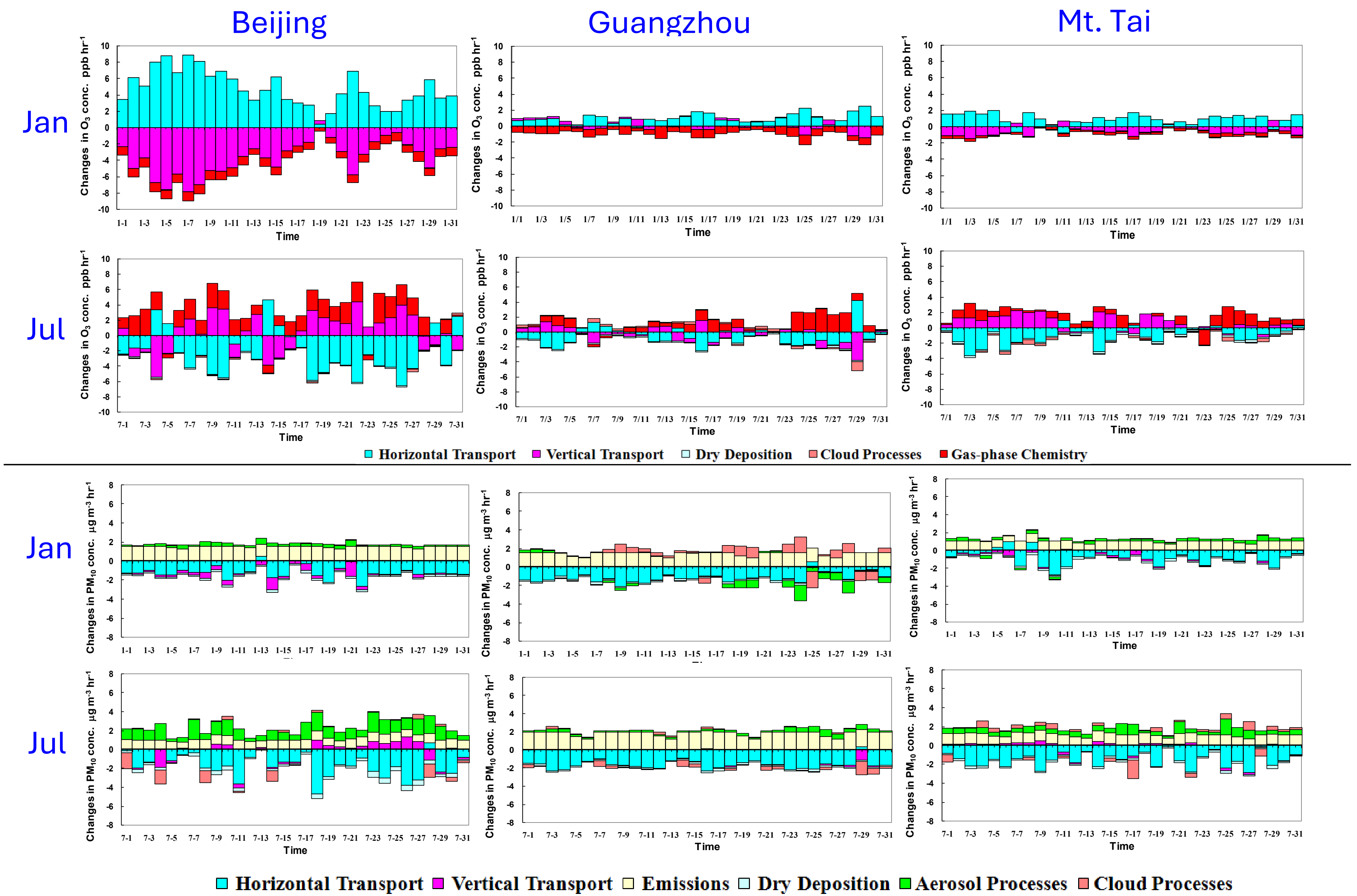
(Zhang et al., 2006)



Default floor value of vertical eddy diffusivity, $K_{zz, min}$ = 1.0,

Sensitivity simulation, $K_{zz, min}$ = 0.1 $m^2 s^{-1}$

Diagnostic Evaluation: Process Analysis of O_3 (top) and $PM_{2.5}$ (bottom) (Liu et al., 2010)



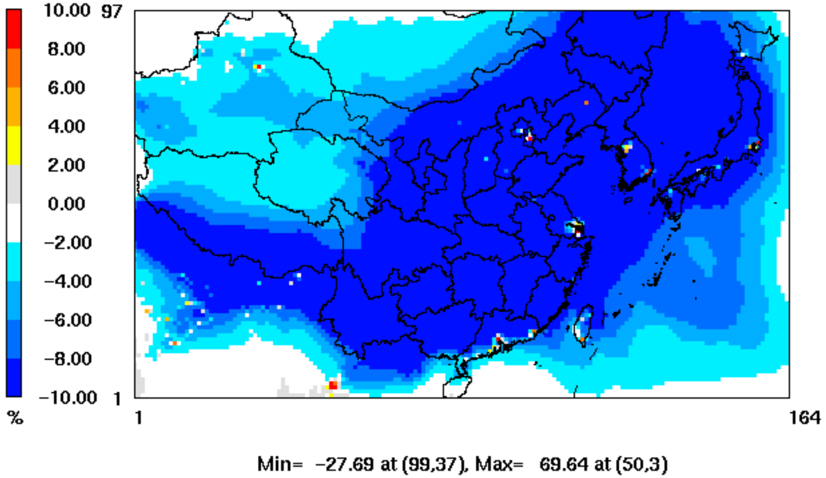
Mechanistic (Dynamic) Evaluation (Zhang, 2024)

- Simulation of several episodes: model responses to meteorology
- Simulation of different areas: model responses to various emission mixtures
- Simulation of different time periods: model responses to changes in emissions (e.g., weekday vs. weekend)
- Simulation under different emission scenarios: NO_x- vs. VOC-limited O_3 chemistry
- Simulation of different emission sectors/areas: source appointment

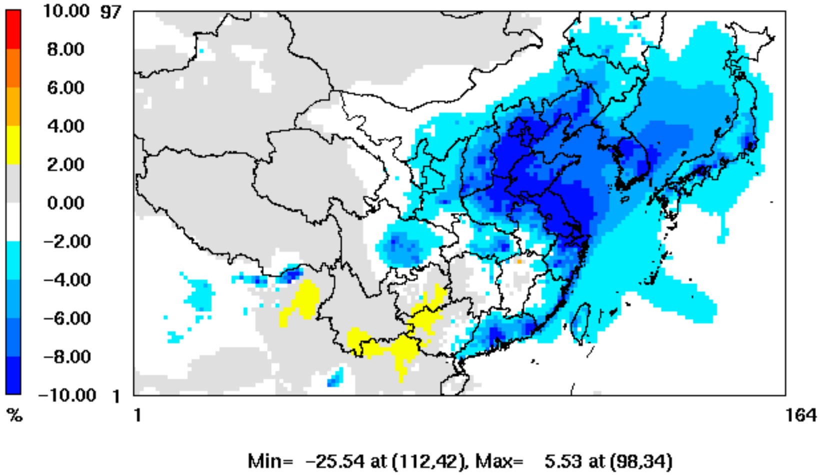
NO_x- vs. VOC limited O_3 Chemistry in China in 2008 (Liu et al., 2010)

Changes in simulated O_3 mixing ratios in Jul, 2008

50% reduction in NO_x emissions



50% reduction in VOCs emissions

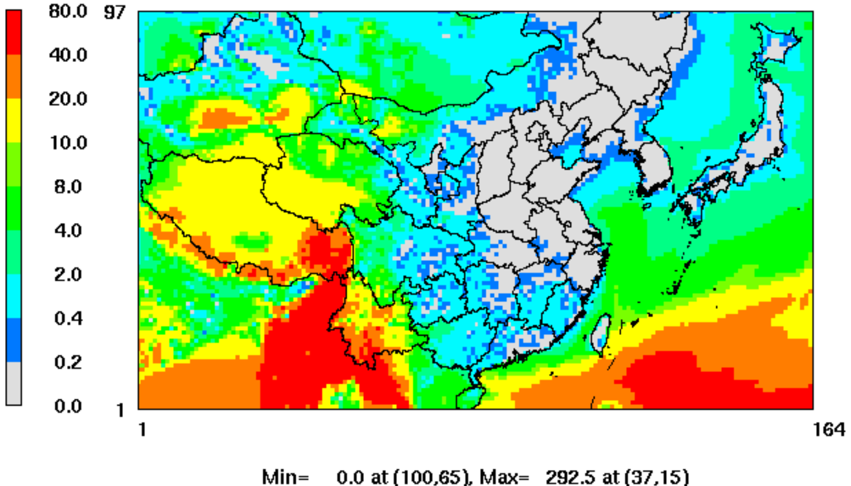


Photochemical indicator $PH_2O_2/PHNO_3$

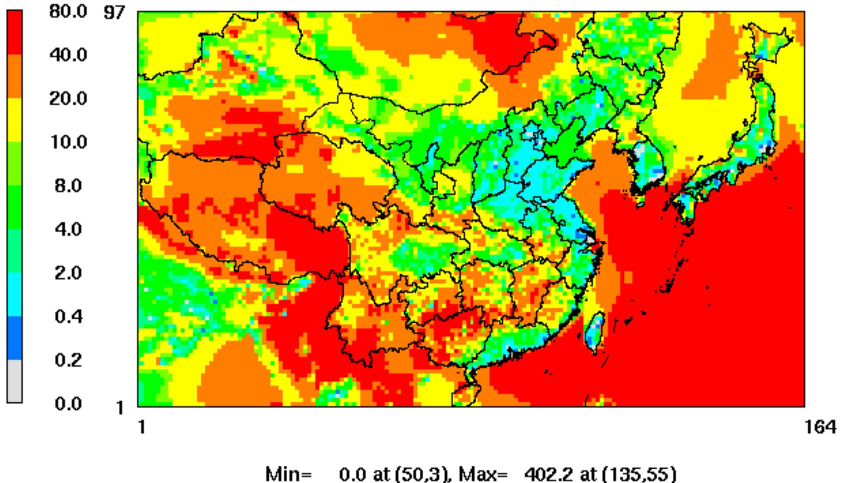
< 0.2, VOC-limited chemistry

≥ 0.2, NO_x-limited chemistry

Jan.

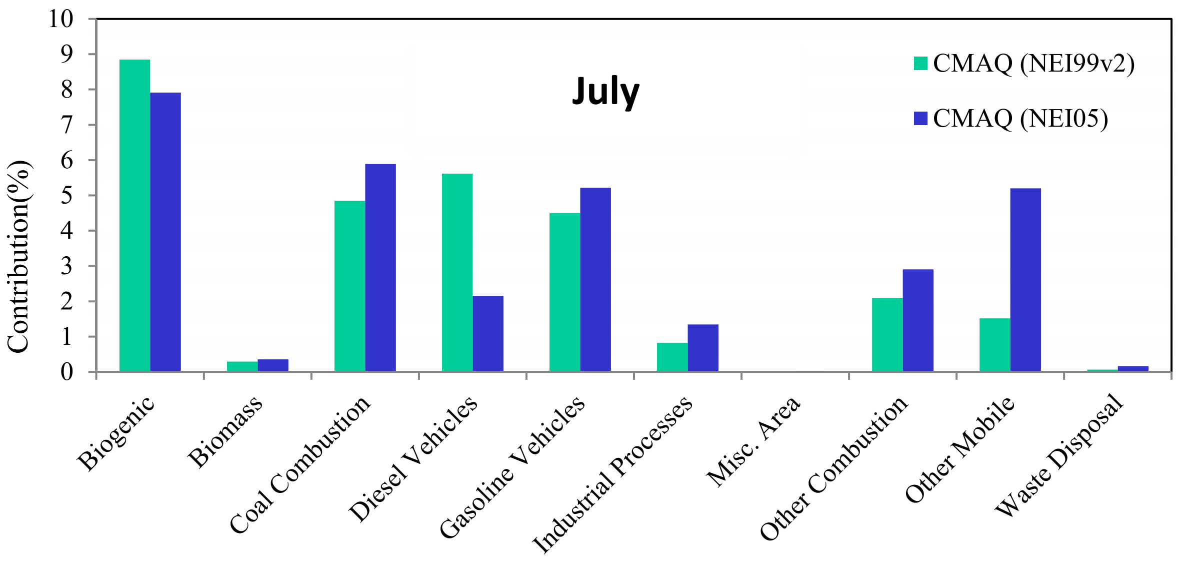


Jul.

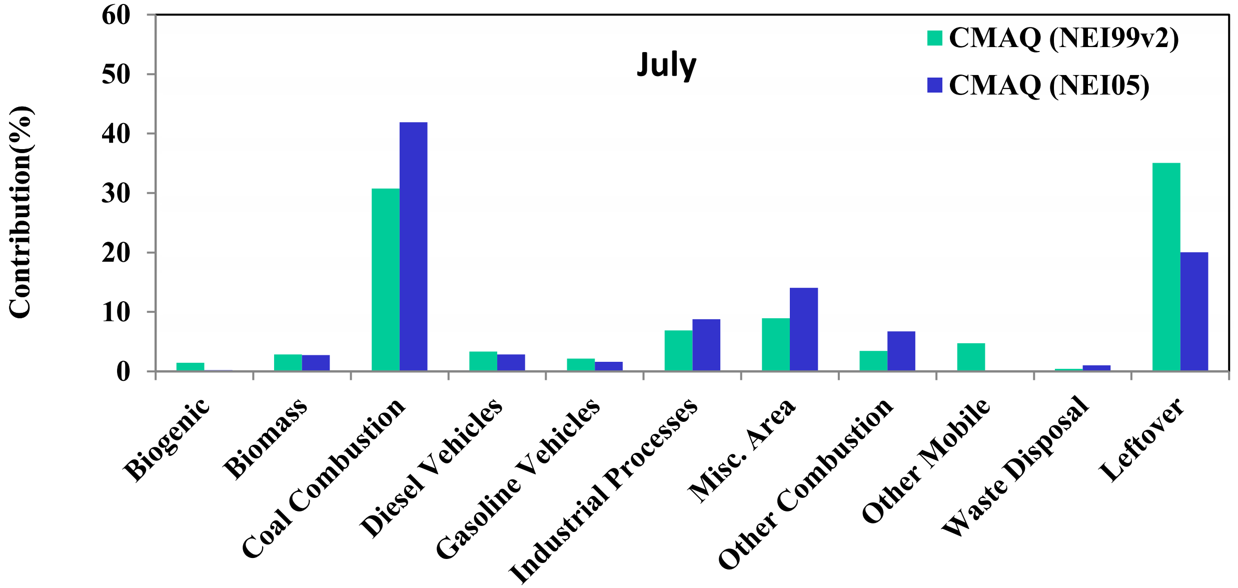


Source Contributions to O_3 and $PM_{2.5}$ over SE U.S. in July 2002 (Burr and Zhang, 2011)

Max 8-h average O_3

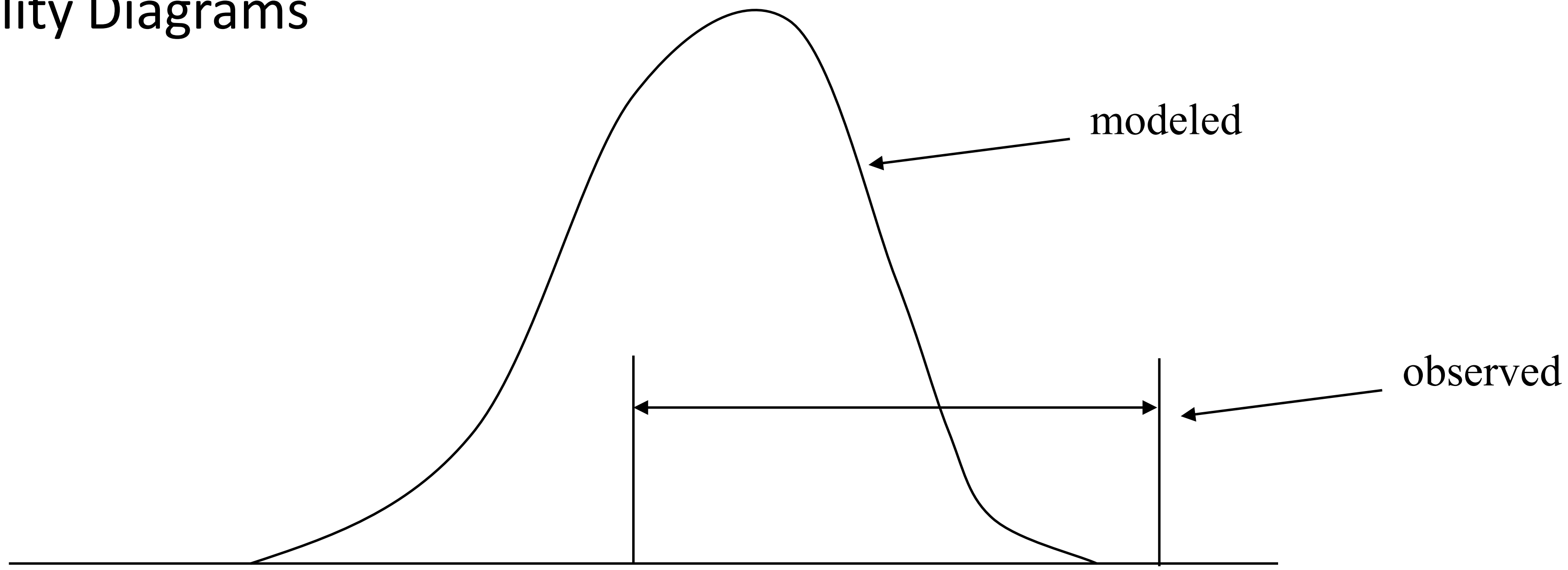


24-hr average $PM_{2.5}$



Probabilistic Evaluation (Zhang, 2024)

- Probability distribution functions (PDFs) for uncertainty and variability of model inputs
- PDFs of model output ($PM_{2.5}$) compared with probability distribution or confidence intervals of observations
- Possible approach to quantify model uncertainty: ensemble modeling with different model configurations or model inputs
  - Talagrand Diagrams (Rank Histograms)
  - Reliability Diagrams



Current Status of Model Evaluation **(Zhang, 2024)**

- Operational evaluation for meteorology and air quality has been extensively performed; increasing numbers of diagnostic and process analysis as well as mechanistic evaluation have been performed; probabilistic evaluation has been less frequently performed but is gaining increasing attentions.
- Various testbeds in the U.S. and in other countries (e.g., Canada, Europe); Community testbeds established by multi-organizations (e.g., the Aerosol InterComparison project (AeroCom), the Air Quality Model Evaluation International Initiative (AQMEII))
- Good performance for O_3 and $PM_{2.5}$ mass concentrations. Relatively poor performance for nitrate and organic aerosols. The performance evaluations, however, are mostly operational. Large uncertainties in predictions of PM number conc. and size distribution.
- Relatively good understanding of oxidant chemistry, but limited understanding of $PM_{2.5}$, particularly organic PM.
- Large uncertainties in model predictions of radiative properties and total column mass conc., due mainly to uncertainties in model treatments of aerosol/cloud microphysics.
- Uncertainties in model inputs (emissions, meteorology, boundary conditions) limit model accuracy, and corroborative modeling techniques have developed and applied to verify model results.

Outline

- **Importance**
- **Model Evaluation**
  - Datasets for Model Evaluation
  - Types of Model Evaluation
  - Current Status of Model Evaluation
- **Preliminary Evaluation of Initial WRF-Chem Application over Africa**
  - Specific Datasets Used for Model Evaluation
  - Evaluation Metrics and Protocols
  - Meteorological Evaluation (WRF only, Jan-April 2023)
  - Chemical Evaluation (WRF-Chem, Jan, 2023)
- **Summary**

Major sources: Zhang et al., 2006a,b, 2019, Zhang (2024), Dennis et al., 2010; Emery and Tai, 2001; Emery et al., 2017

Data for Model Evaluation in Africa

Africa Air monitoring networks

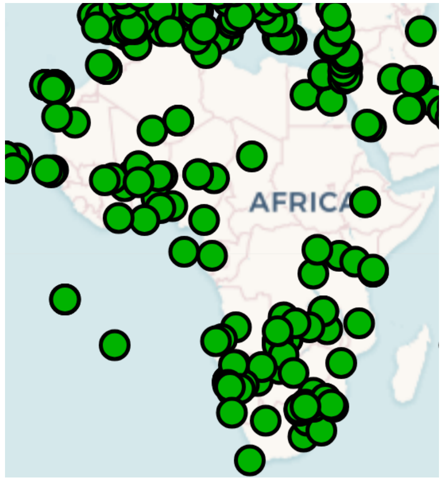
Network	Region	Total site #	Variable measured	Temporal resolution	Measurement method
SAAQIS	South Africa	175	CO, NO, NO_2, SO_2, O_3, $PM_{2.5}$, PM_{10}, meteorology	Hourly / - Present	Research grade
RBCAA	South Africa	10	$PM_{2.5}$, PM_{10}, SO_2, meteorology	Hourly / - Present	Research grade
AirNow	Egypt	1	$PM_{2.5}$	Hourly/2022 - Present	FEM & low-cost
EEAA	Egypt	120	PM_{10}, NO, NO_2, SO_2, O_3, CO	1998 – Present	Research grade
CAIP	Egypt	37	$PM_{2.5}$, PM_{10}	1998 – 2007	Research grade
AfriqAir	Ivory Coast	2	$PM_{2.5}$, NO_x, O_3	Hourly/ - Present	FEM&low-cost
AirQo	Africa	>250+	$PM_{2.5}$, PM_{10}	Hourly/ 2023-Present	Low-cost sensor
EMA	Egypt	1	PM_{10}	Hourly/- Present	Research grade
ISD	Worldwide	20,000	Meteorology	Hourly/1929 - Present	Research grade

Ground AOD

Products	Region	Species	Temporal resolution/time period
AERONET	Worldwide	AOD	Daily and Monthly / 2014 – Present

Satellite AOD

Products	Region	Resolution	Species	Temporal resolution/time period
MODIS	Worldwide	1-km	AOD	Daily / 2014 – Present
CERES-MODIS	Worldwide	1°x1°	SWR	Daily / - Present
CLARA	Worldwide	0.25°x0.25°	Surface radiation budget	Daily / 1979 - Present



Reanalysis MERRA-2

Model	Region	Resolution	Species	Temporal resolution/time period
MERRA 2	Worldwide	0.625°x0.5°	Surface: BC, dust, OC, $PM_{2.5}$, sulfate, sea salt, SO_2 Total Column: BC, dust, OC, O_3, sulfate, sea salt, SO_2	Hourly / - Present

aeronet.gsfc.nasa.gov

Data used for the African Testbed Evaluation

- Meteorology – NOAA Global Hourly Integrated Surface Database
  - 20,000 stations worldwide
  - Data includes wind speed (WS), wind direction (WD), temperature (T), and dew point temperature (DT)
- Air quality
  - **AirNow**: ground observations for $PM_{2.5}$
  - **EMA:** ground observationPM_{10} data
  - **AERONET**: ground truth for AOD
  - **MERRA-2**: reanalysis data for $PM_{2.5}$

Model Evaluation Metrics and Protocols

Meteorological variables	MB	IOA	RMSE	NMB	References
T2	≤ |±0.5|	≥ 0.8	——	——	Emery et al. (2001)
WS10	≤ |±0.5|	≥ 0.6	≤ |±2|	——	Emery et al. (2001)
WD10	≤ |±10|	——	——	——	Zhang et al. (2006a, 2019)
Precipitation	——	——	——	< |±30%|	Zhang et al. (2006a, 2019)

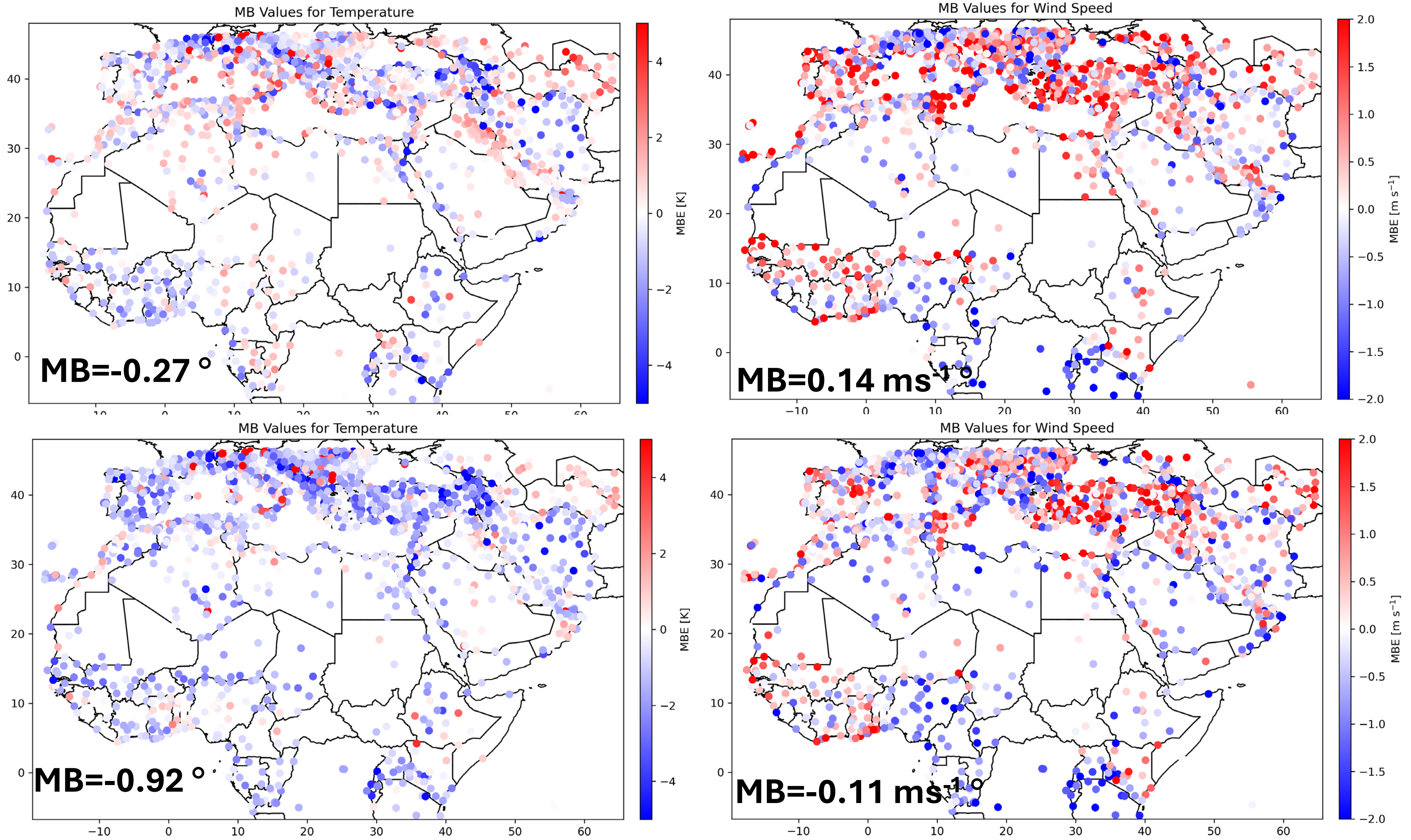
Air pollutant variables	NMB		NME		R		
	Goal	Criteria	Goal	Criteria	Goal	Criteria	
Max 8h O_3	< |±5%|	< |±15%|	< |15%|	< |25%|	> 0.75	> 0.5	Emery et al. (2017), Zhang et al. (2006b)
24-hr SO_4^{2-}, NH_4^+, $PM_{2.5}$	< |±10%|	< |±30%|	< |35%|	< |50%|	> 0.7	> 0.4	
24-hr NO_3^-	< |±15%|	< |±65%|	< |65%|	< |115%|	——		
24-hr OC	< |±15%|	< |±50%|	< |45%|	< |65%|	——		
24-hr EC	< |±20%|	< |±40%|	< |50%|	< |75%|	——		

- Good model performance falls within the range of the benchmark values
- Benchmarks vary by regions
- Newer models tend to produce higher benchmark scores
- Seasonal variations result different metrics

WRF Spatial Evaluation (WRF only)

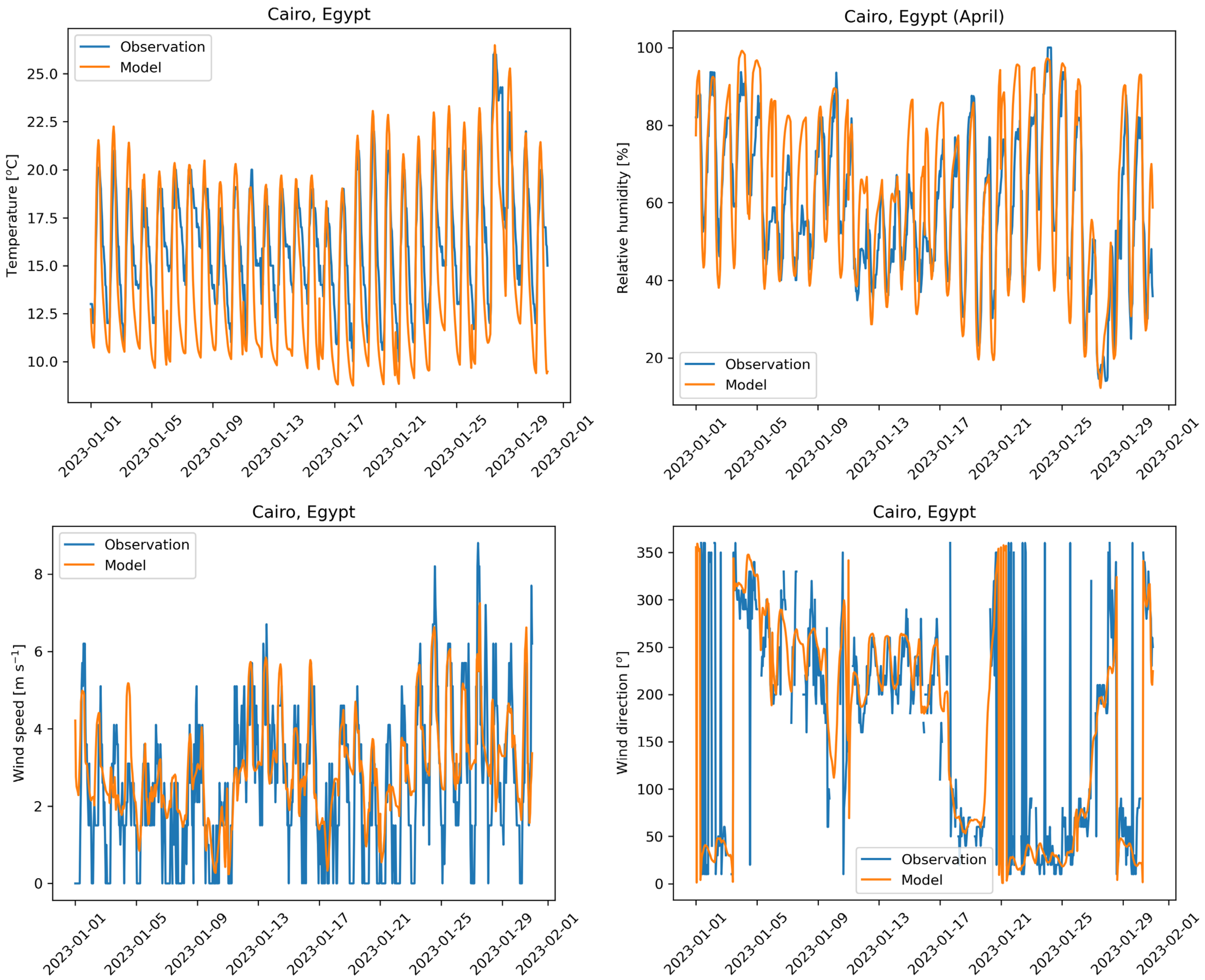
January

April



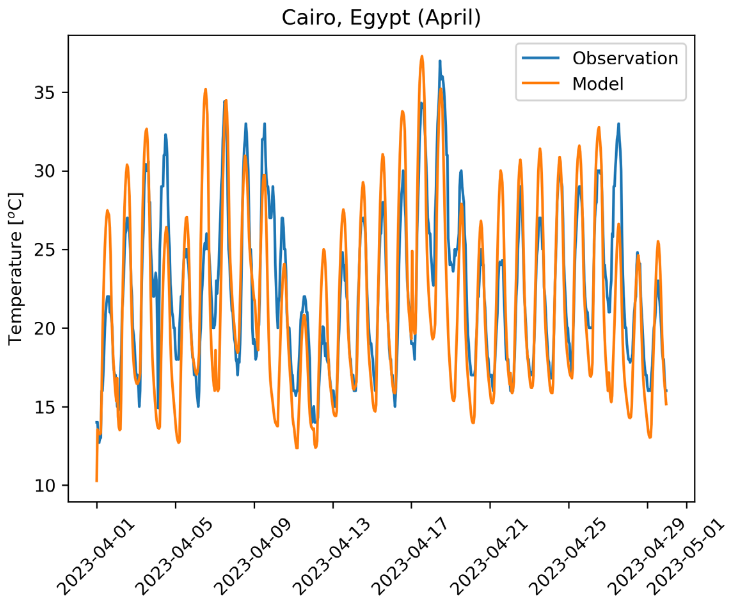
- T2 is well predicted in January but gives a cold bias in April
- WS10 is overpredicts in January but underpredicts in April

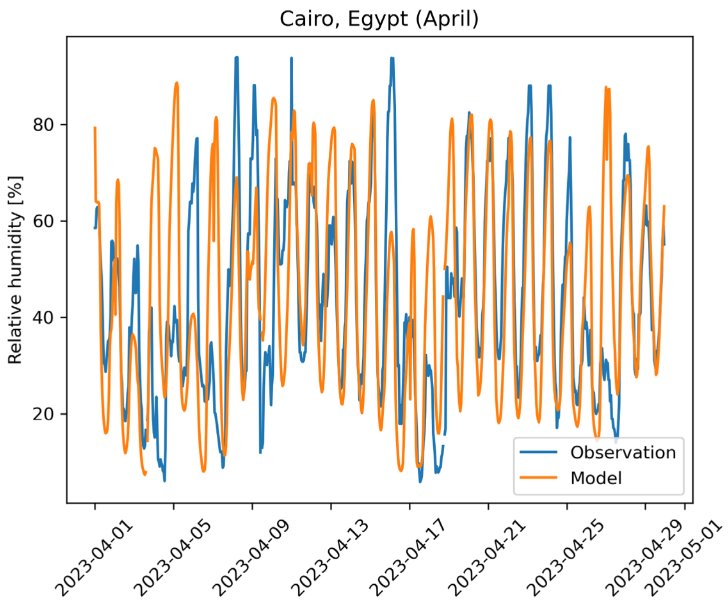
January Timeseries - Cairo, Egypt (WRF only)

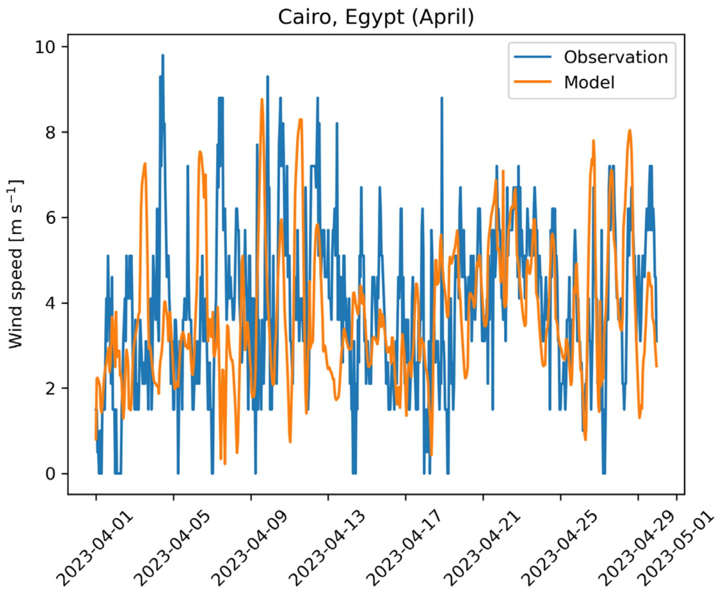


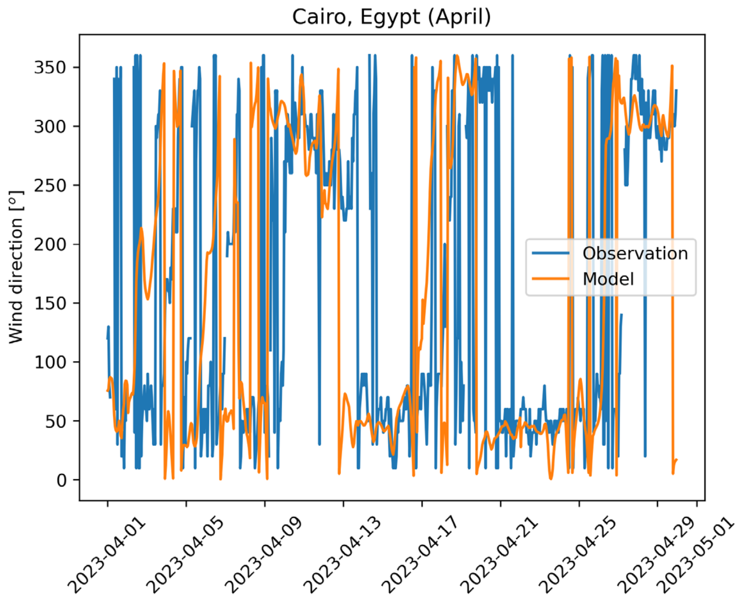
- WRF exhibits cold bias in temperature predictions for Cairo
- The model shows good correlation with RH
- WRF underpredicts wind speed, but has good correlation with wind direction

April Timeseries - Cairo, Egypt (WRF only)









- WRF captures temperature trends and RH in Cairo during April
- Good performance on wind speed

$PM_{2.5}$ Evaluation

- Time conversion: local time to UTC
- Monitoring networks: no conversion is needed for ground observations
- MERRA-2: $PM_{2.5}$ can be calculated using Buchard et. el., 2016

$$PM_{2.5} = PM_{2.5}^{DU} + PM_{2.5}^{SS} + PM_{2.5}^{OC} + PM_{2.5}^{BC} + \left(\frac{132.14}{96.06}\right) PM_{2.5}{}_{4}^{SO_4}$$

  - $PM_{2.5}^{DU}$ is dust, $PM_{2.5}^{SS}$ is sea salt, $PM_{2.5}^{OC}$ is organic carbon, $PM_{2.5}^{BC}$ is black carbon, and $PM_{2.5}{}_{4}^{SO_4}$ is sulfate

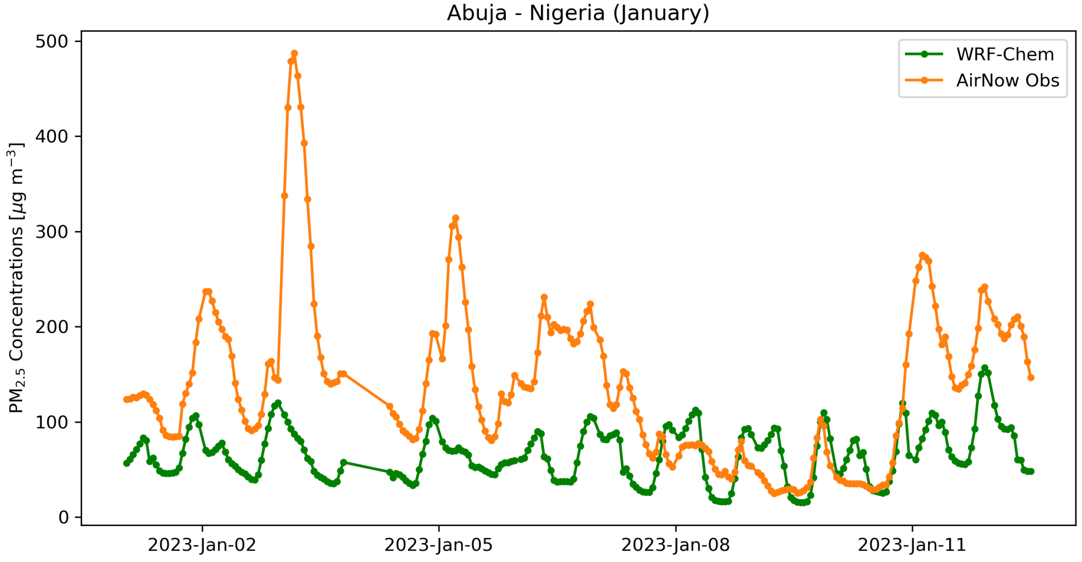
- AOD evaluation
  - AOD from AERONET
    - Interpolate AOD 550nm from AERONET dataset
  - AOD from WRF-Chem

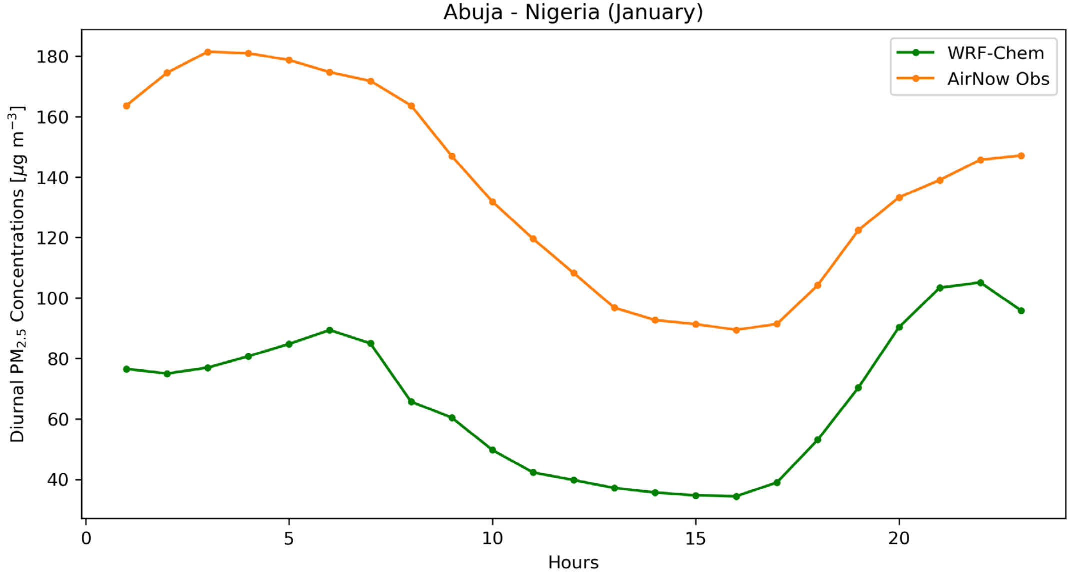
$$\text{WRF } AOD550 = \sum_{n=1}^{N} EXTCOF55 * Z$$

where $Z = \frac{PH+PHB}{9.8}$

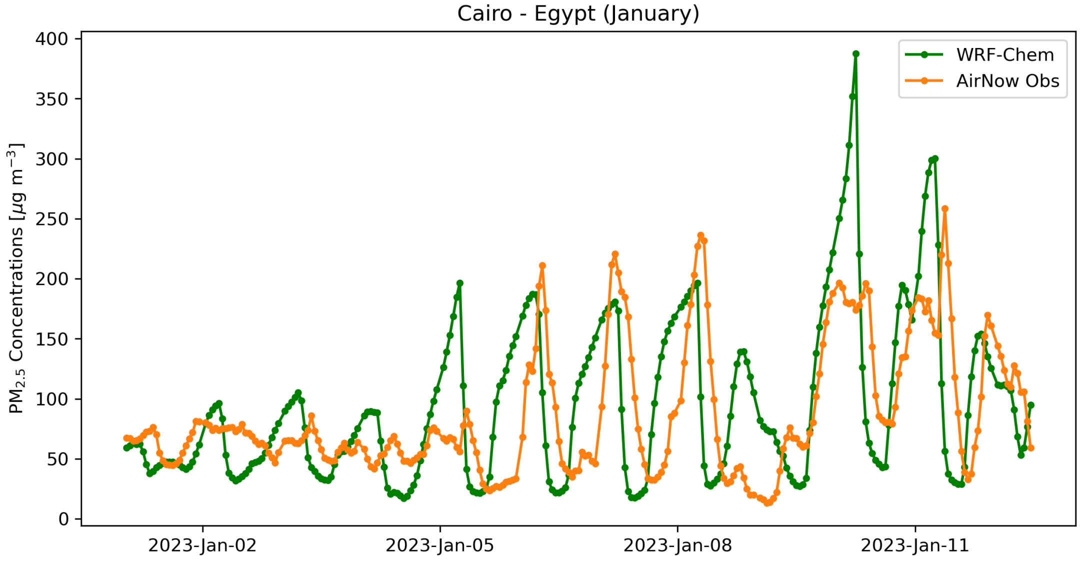
PM$_{2.5}$ Timeseries (January)

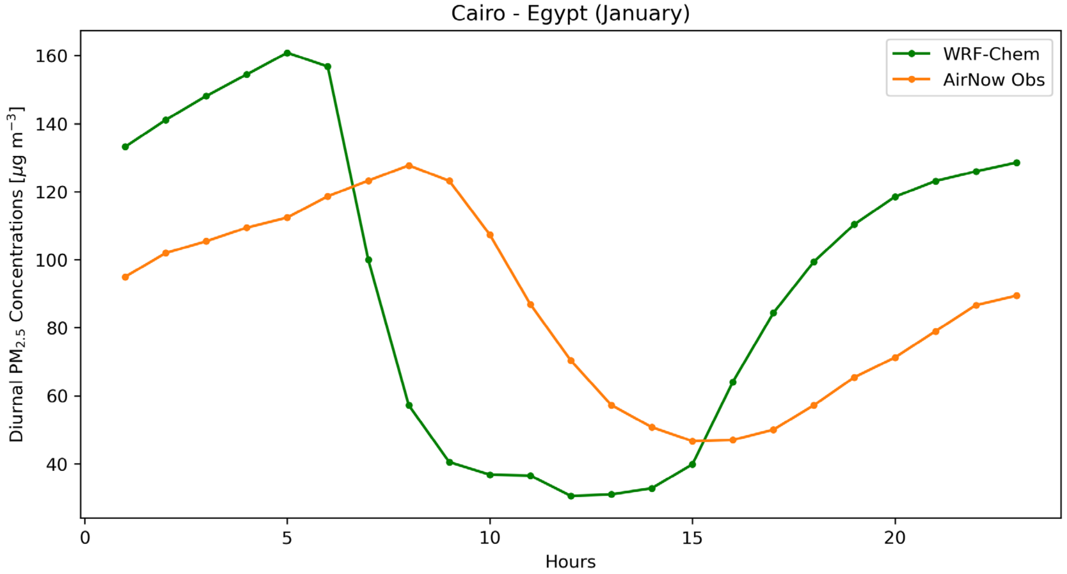
Abuja, Nigeria



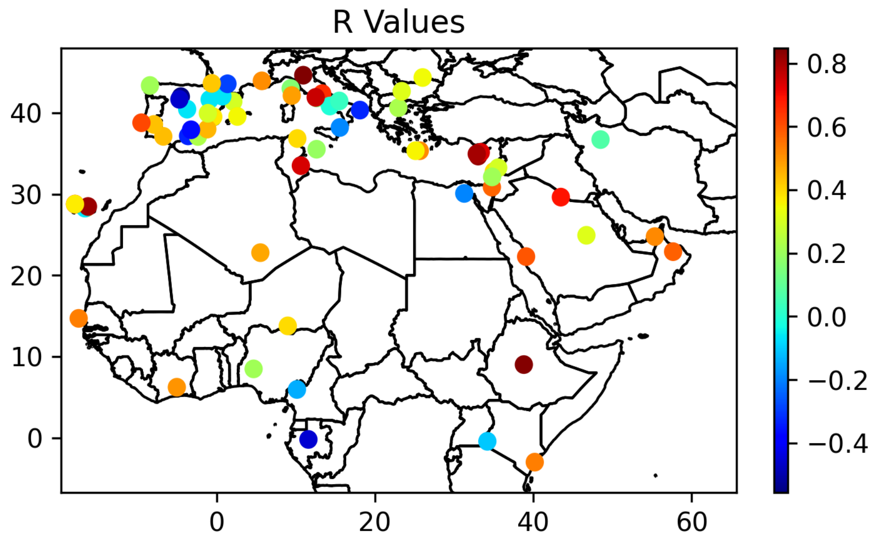


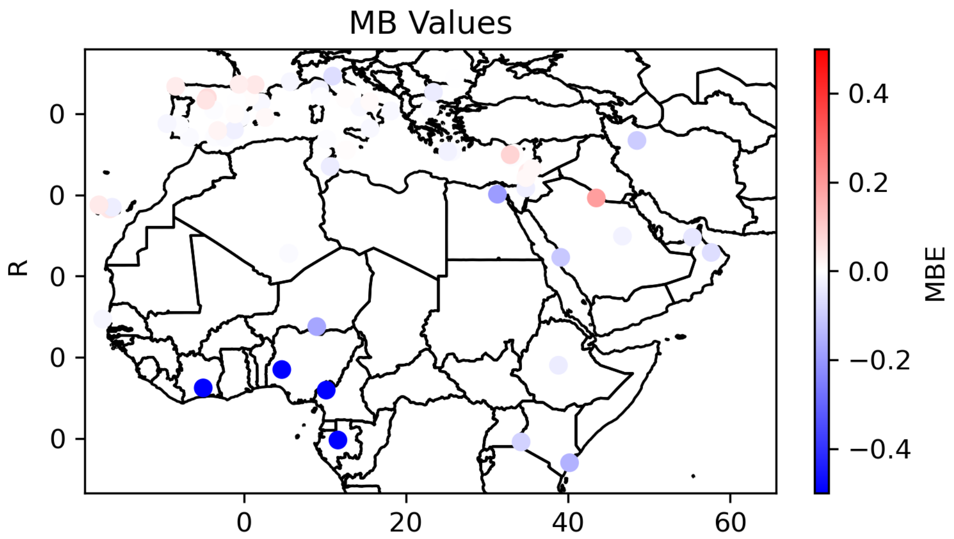
Cairo, Egypt

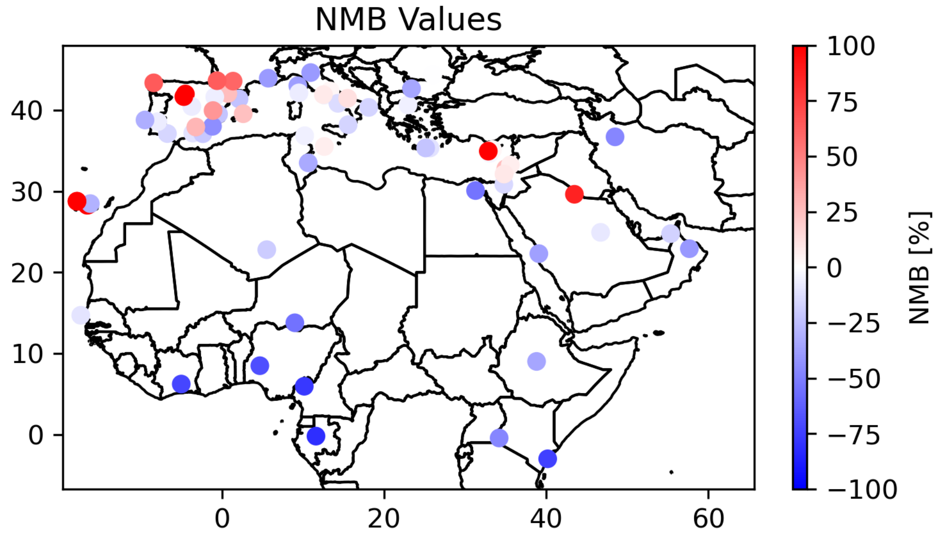


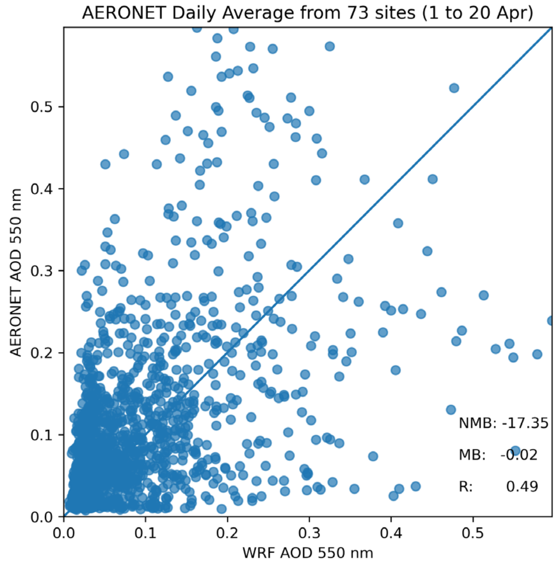


AOD Evaluation Against AERONET (January)





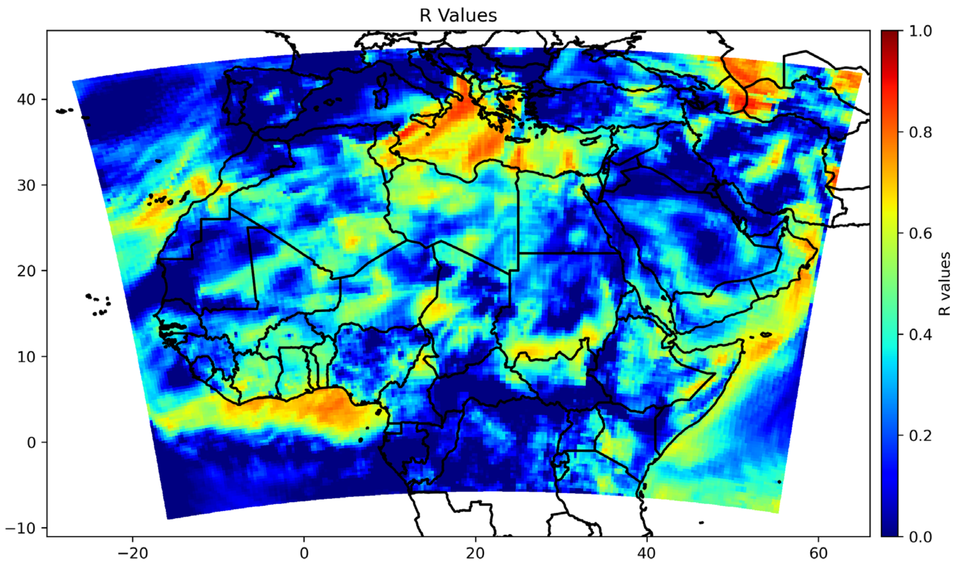


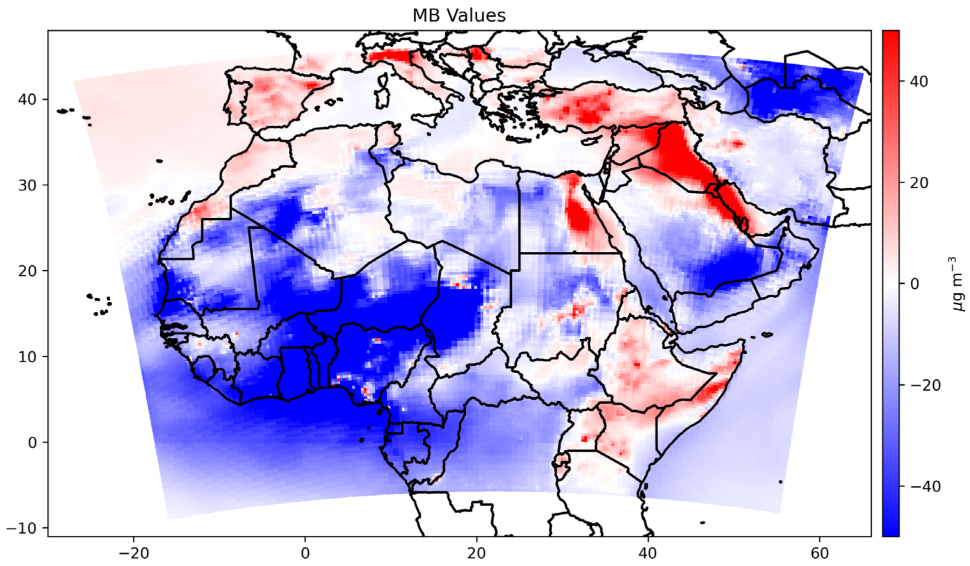


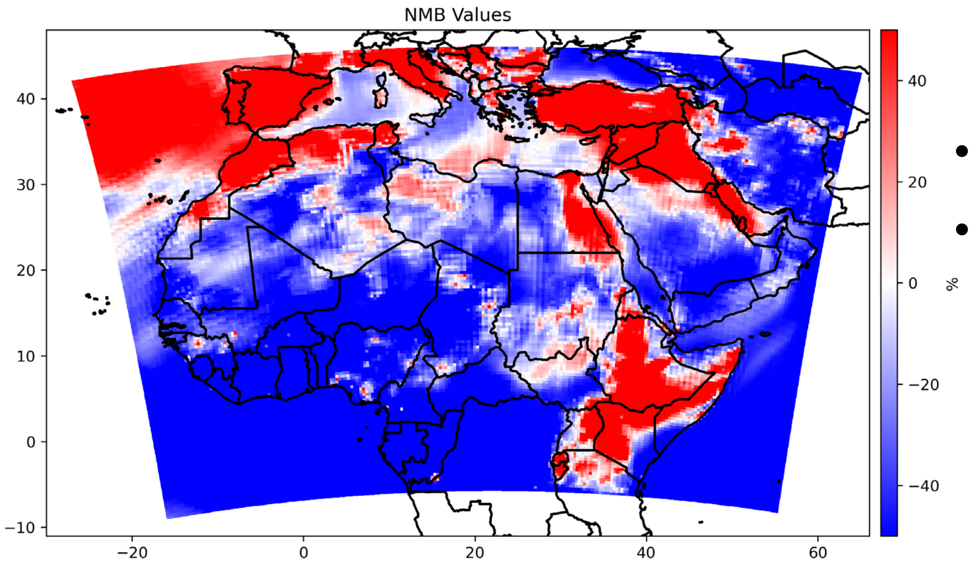
R	MBE	NMB
0.49	-0.02	-17.35

- Reasonably good correlation between predicted and observed AOD
- WRF-Chem shows moderate underprediction in AOD, due likely to the underpredictions in PM

$PM_{2.5}$ Spatial Evaluation Against MERRA-2 (January)



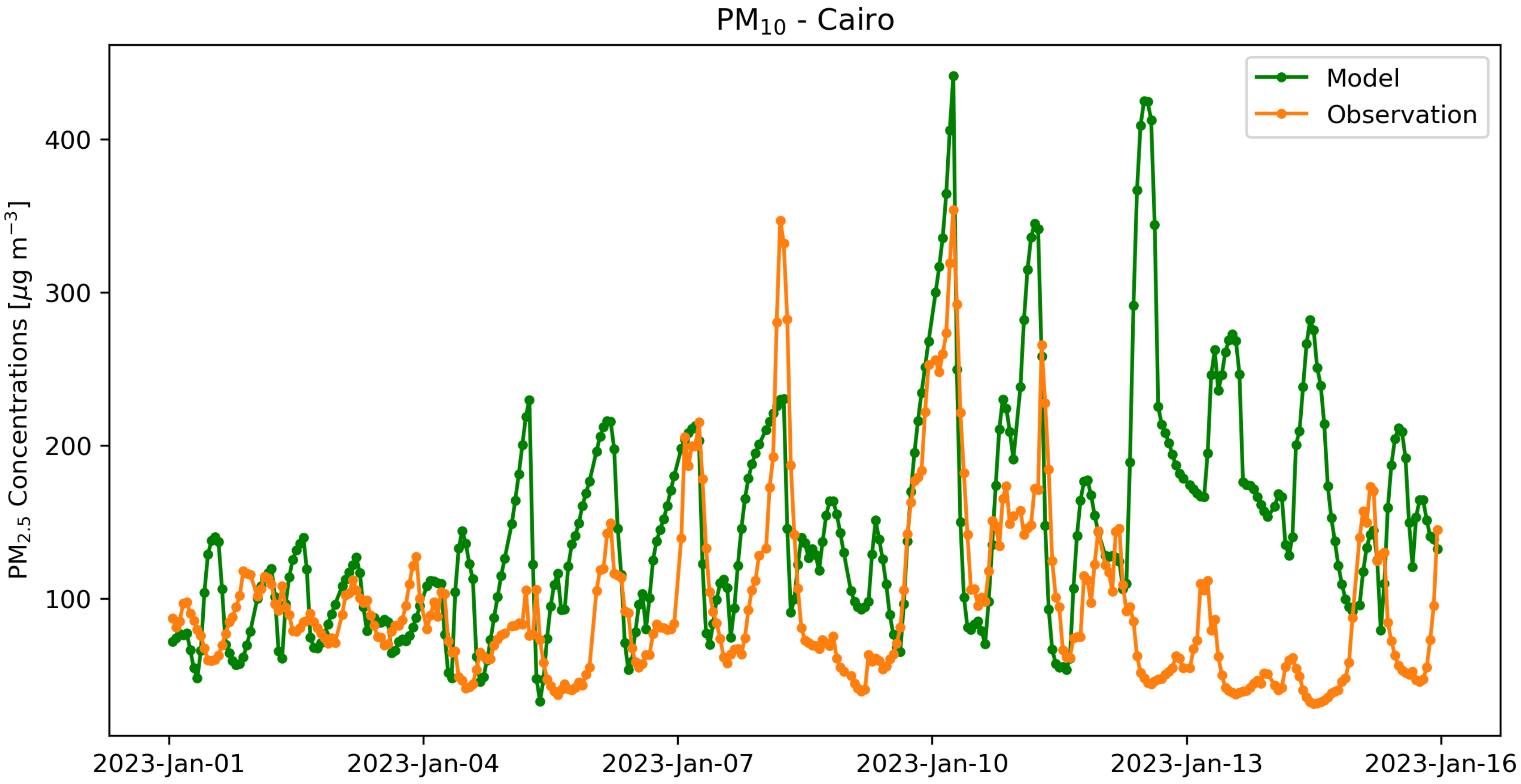




- WRF-Chem shows a good agreement in terms of NMB
- Low MB in the North and East Africa regions

R	MB	NMB%
0.26	-9.27	-2.33

PM_{10} Evaluation at the EMA site in Cairo (January)



Statistics

R = 0.26

MB = 45 $\mu g\ m^{-3}$

NMB = 48%

- Reasonably good agreement during Jan 1-12
- Largely PM_{10} overpredictions during Jan 13-15

Summary

- Model evaluation is a critical step to establish model fidelity to support decision-making and create creditable baseline for future projection.
- Increasing number of datasets are available for model evaluation in many regions, including Africa, each having its own merits. Only calibrated data should be used, and QA/QC is critical to ensure the data quality.
- Major types of model evaluation include operational, diagnostic, mechanistic (aka dynamic), and probabilistic, offering complementary information to comprehensively assess the model's skill and associated sensitivity and uncertainties.
- Preliminary evaluation of the initial application of WRF and WRF-Chem in Africa shows some skills but more work remain to identify sources of errors and improve the performance

References

- Dennis, R., et al. (2010). A framework for evaluating regional-scale numerical photochemical modeling systems. J. Environ. Fluid Mech., 10:471–89.
- Emery, C., and E. Tai (2001), Enhanced Meteorological Modeling and Performance Evaluation for Two Texas Ozone Episodes, project report prepared for the Texas Natural Resource Conservation Commission, Prepared by ENVIRON, International Corporation, Novato, CA.
- Emery, C., Z. Liu, A.G. Russell, M.T. Odman, G. Yarwood, and N. Kumar (2017), Recommendations on statistics and benchmarks to assess photochemical model performance, J. Air Waste Manage. Assoc., DOI:10.1080/10962247.2016.1265027: http://dx.doi.org/10.1080/10962247.2016.1265027.
- Zhang, Y., P. Liu, B. Pun, and C. Seigneur, 2006a, A Comprehensive Performance Evaluation of MM5-CMAQ for the Summer 1999 Southern Oxidants Study Episode, Part-I. Evaluation Protocols, Databases and Meteorological Predictions, *Atmospheric Environment*, 40, 4825-4838, doi: 10.1016/j.atmosenv.2005.12.043.
- Zhang, Y., P. Liu, A. Queen, C. Misenis, B. Pun, C. Seigneur, and S.-Y. Wu, 2006b, A Comprehensive Performance Evaluation of MM5-CMAQ for the Summer 1999 Southern Oxidants Study Episode, Part-II. Gas and Aerosol Predictions, *Atmospheric Environment*, 40, 4839-4855, doi:10.1016/j.atmonsenv.2005.12.048.
- Zhang, Y., C. Jena, K. Wang, C. Paton-Walsh, E.-A. Guerette, S. Utembe, J. D. Silver, and M. Keywood, 2019, Multiscale Applications of Two Online-Coupled Meteorology-Chemistry Models during Recent Field Campaigns in Australia, Part I: Model Description and WRF/Chem-ROMS Evaluation Using Surface and Satellite Data and Sensitivity to Spatial Grid Resolutions, Atmosphere, *10(4)*, 189, 1-44; doi:10.3390/atmos10040189.
- Zhang, Y., 2024, Air Quality in a Changing Climate: Science and Modeling, Cambridge University Press, in preparation.