Dear Dr. Juan A. Añel,

Thank you for your email regarding the GMD Code and Data Policy.

We wish to clarify that the ChinaHighPMC dataset was not used in this study. The concern regarding its repository status is therefore not applicable to our manuscript.

All input data employed in our work are explicitly documented in Section 2.2 (“Data”) and are publicly available from the following repositories:

Datasets	Download DOI or website	Paper DOI
CAQRA	https://doi.org/10.11922/sciencedb.00053	https://doi.org/10.5194/essd-13-529-2021
CAQRA-aerosol	https://doi.org/10.12423/capdb_PKU.2023.DA	https://doi.org/10.1007/s00376-024-4046-5
TAP	http://tapdata.org.cn	https://doi.org/10.1021/acs.est.2c06510
NP2	https://doi.org/10.5281/zenodo.10886914	https://doi.org/10.5194/gmd-17-8495-2024
CAMSRA	https://ads.atmosphere.copernicus.eu/datasets/cams-global-reanalysis-eac4?tab=overview	https://doi.org/10.5194/acp-19-3515-2019
MERRA-2	https://disc.gsfc.nasa.gov/datasets?project=MERRA-2	https://doi.org/10.1175/JCLI-D-16-0609.1

Furthermore, all output data from our experiments, including the test data, assimilation system outputs, and the reanalysis data used for comparison, have been publicly deposited on ZENODO (https://doi.org/10.5281/zenodo.16735735). This has been clearly declared in the "Code and Data Availability" section of our manuscript.

All data and code related to this study are thus publicly accessible and fully compliant with GMD's policy.

For your convenience, we have attached the version of the manuscript that passed the “files validated” stage on 15 August 2025. We trust this clarifies the matter and confirms our full adherence to the journal's policy.

Should you require any further clarification, we are available to provide it.

Sincerely,
Ting Yang
Institute of Atmospheric Physics, Chinese Academy of Sciences

OIRF-LEnKF v1.0: A Self-evolving Data Assimilation System by Integrating Incremental Machine Learning with a Localized EnKF for Enhanced $PM_{2.5}$ Chemical Component Forecasting and Analysis

Hongyi Li[1], Ting Yang[1*], Lei Kong[1], Di Zhang[2], Guigang Tang[2], Zifa Wang[1, 3]

[1]State Key Laboratory of Atmospheric Environment and Extreme Meteorology, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029, China.
[2]China National Environmental Monitoring Centre, Beijing, China
[34]College of Earth and Planetary Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

Correspondence to: Ting Yang (tingyang@mail.iap.ac.cn)

Abstract. Assimilating observational data into numerical forecasts is crucial for accurately estimating the spatiotemporal distribution of $PM_{2.5}$ chemical components (NH_4^+, NO_3^-, SO_4^{2-}, OC, and BC), which is beneficial to quantifying the impact of aerosols on the environment, climate change and human health. However, chemical transport model (CTM)-based data assimilation (DA) is computationally inefficient for large ensemble sizes and offers limited improvements in forecasting, as it solely provides optimal initial conditions. This paper introduces a machine learning (ML)-based self-evolving data assimilation system (OIRF-LEnKF v1.0) that achieves high efficiency and high quality in the forecast and analysis fields of chemical components. Computational efficiency tests indicate that the total time consumed by OIRF-LEnKF v1.0 constitutes only 11.41-16.60 % of that of CTM-based DA, particularly during the forecasting process (0.13-0.20 %). Sensitivity tests demonstrate that the self-evolution mechanism in our system enhances the Pearson correlation coefficient (CORR) and reduces the RMSE during the forecasting process by 2.28-11.75 % and 32.94-40.98 %, respectively, compared to the stationary training mechanism. A 2-month DA experiment reveals that the RMSE values of chemical components after DA are less than 7.80 μg m^{-3} and 2.36 μg m^{-3} during the forecasting and analysis processes, respectively, indicating reductions of at least 26.38 % and 68.99 % compared to values without DA. Notably, the RMSE values of our system during the forecasting process exhibit a significant reduction of 33.16-90.10 % compared to those of the CTM-based DA, highlighting the superior forecasting capability of our system. Furthermore, the spatial overestimation and underestimation of chemical components have been significantly mitigated following DA. Compared to multiple reanalysis datasets of inorganic salt aerosols (CORR: 0.56-0.89, RMSE: 2.55-8.52 μg m^{-3}), the dataset generated by OIRF-LEnKF v1.0 (CORR: 0.97, RMSE: 1.12 μg m^{-3}) demonstrates higher data quality.

1 Introduction

Sulfate (SO_4^{2-}), nitrate (NO_3^-), ammonium (NH_4^+), organic carbon (OC), and black carbon (BC) are critical chemical components of fine particulate matter ($PM_{2.5}$) (Huang et al., 2014). The physicochemical processes of these chemical components within the atmospheric boundary layer, including chemical conversion, transboundary transport and deposition,

directly influence air quality associated with $PM_{2.5}$ (Yang et al., 2024). Observational studies reveal that the contribution of transboundary transport increased from 4-8 % to 66-80 % during severe $PM_{2.5}$ pollution episodes (Sun et al., 2016). Furthermore, these components with varying physicochemical properties exert varying impacts on human health (Li et al., 2022) and climate change (Stier et al., 2024; Zhao et al., 2024). Therefore, characterizing the spatiotemporal distribution and evolution of $PM_{2.5}$ chemical components provides a scientific basis for identifying the causes of air pollution, assessing health and climate impacts, and developing effective climate change mitigation strategies and emission pathways.

Observation techniques, machine learning (ML) methods, and chemical transport models (CTMs) are the primary approaches for acquiring mass concentrations of $PM_{2.5}$ chemical components. Observation techniques achieve high-precision measurements through field sampling and instrument analysis (Wang et al., 2016; Lei et al., 2021). However, the sparse distribution of observation points, limited observation pathways, inconsistencies in observation platforms, and measurement errors hinder the acquisition of continuous measurements with high spatiotemporal coverage. ML methods utilize historical observations to establish mapping relationships between features of non-chemical and chemical components, thereby reconstructing the mass concentrations of chemical components continuously without the need for traditional instrument measurements (Li et al., 2025; Wei et al., 2023; Liu et al., 2022). However, ML methods are limited by the lack of physicochemical constraints and insufficient spatiotemporal representativeness of historical observations, which results in inadequate generalization capabilities and interpretability. CTMs can characterize the spatiotemporal distribution and evolution of chemical components by solving equations that describe physicochemical mechanisms rather than relying on observations (Weagle et al., 2018). However, the uncertainties in physicochemical mechanisms, emission inventories, meteorological fields, as well as initial and boundary conditions result in significant simulation bias (Miao et al., 2020; Xie et al., 2022; Luo et al., 2023).

Data assimilation (DA) can integrate observations from sparse sites and CTMs to estimate an optimal initial state with spatial continuity and high accuracy based on the model forecast field (Geer, 2021). DA technique has been widely used to generate reanalysis datasets of $PM_{2.5}$ chemical components at global and national scales, such as the Copernicus Atmosphere Monitoring Service ReAnalysis (CAMSRA) (Inness et al., 2019), the Modern-Era Retrospective Analysis for Research and Applications Version 2 (MERRA) (Randles et al., 2017), and the Air Quality ReAnalysis in China dataset (CAQRA-aerosol) (Kong et al., 2025). However, these datasets only assimilate the aerosol optical depth and conventional atmospheric pollutants at the surface level, indirectly enhancing simulations of chemical components. Consequently, the correlation between observations and these datasets is not statistically significant (R: 0.21 to 0.7) (Kong et al., 2025).

Our previous work developed a novel hybrid nonlinear ensemble data assimilation system (NAQPMS-PDAF v2.0, NP2) for directly assimilating observations of chemical components (Li et al., 2024). However, CTM-based NP2 requires a reduction in ensemble size to maintain computational efficiency during forecasting and assimilation processes within high-dimensional

state spaces, resulting in insufficient ensemble spread (Chattopadhyay et al., 2023). Consequently, the correlation (R: 0.12-0.72) between observations and analysis fields at independent validation sites showed only minor improvement compared to the datasets mentioned above. Furthermore, the low sensitivity of forecast fields in NP2 to assimilation frequency suggests that improvements in initial conditions have limited effects on enhancing forecast ability on $PM_{2.5}$ chemical components due to the uncertainties in physicochemical mechanisms and input conditions within CTMs (Cha et al., 2025).

In recent years, the combination of ML and DA has emerged as a pivotal strategy for addressing challenges associated with computational inefficiency and insufficient improvements in forecasting and analysis fields. The first pathway employs the ML outputs as external constraints for DA, such as forecasting addition (Lin et al., 2019; Jin et al., 2019), bias correction (Arcucci et al., 2021; Farchi et al., 2021; He et al., 2023), parameter estimation (Legler and Janjić, 2022), and observation operator improvement (Lee et al., 2022). This pathway enhances forecasting and DA processes without perturbing the physical properties of the numerical models but fails to improve computational efficiency. The second pathway utilizes ML as an alternative to DA for generating analysis fields directly from high-density observations (Howard et al., 2024). This pathway mitigates the limitations of traditional DA algorithms in handling high-resolution observations while diminishing the physical dependence of observation propagation within model state space. The third pathway substitutes traditional numerical models with ML models to provide the forecast fields for DA (Dong et al., 2022; Dong et al., 2023; Yang and Grooms, 2021) and utilize the analysis fields to update ML model parameters, thereby enhancing forecasting performance (Brajard et al., 2020; Gottwald and Reich, 2021). This pathway improves computational efficiency by 78.3% while maintaining high DA accuracy (Dong et al., 2022) and mitigates the adverse impact of low-quality data on ML forecasting (Buizza et al., 2022). However, to the best of our knowledge, this pathway has not yet been utilized in atmospheric chemical DA.

This study proposes an optimized incremental Random Forest (OIRF) forecasting model as a solution to the challenges of computational inefficiency and inadequate advancements in forecasting and analysis fields within traditional CTM-based DA. The OIRF model is capable of generating a large number of forecasting ensemble members at a reduced computational cost, which mitigates the underestimation of forecast error covariance. Additionally, it can dynamically update by integrating new training data, allowing it to adapt to the evolving dynamics of $PM_{2.5}$ chemical components, thereby enhancing its generalization capability for forecasting. Then, the OIRF model is online coupled with the localized ensemble Kalman filter (LEnKF) algorithm to develop a novel self-evolving data assimilation system (OIRF-LEnKF v1.0), which achieves a rapid iteration for high-quality forecasting, assimilation, and incremental learning. Section 2 details the development of OIRF-LEnKF v1.0, the data used in this study and experimental settings. Section 3 presents the DA results, including an evaluation of computational efficiency, a discussion of sensitivity tests, and a validation of DA performance. Section 4 summarizes the conclusions.

2 Method and Data

2.1 OIRF-LEnKF v1.0

2.1.1 Structure of OIRF-LEnKF v1.0

The OIRF-LEnKF v1.0 performs a continuous loop of forecasting and assimilation for five $PM_{2.5}$ chemical components (SO_4^{2-}, NO_3^-, NH_4^+, OC, and BC) through online coupling an optimized incremental Random Forest (OIRF) ensemble forecasting model with the localized ensemble Kalman filter (LEnKF) algorithm (Fig. 1). The ML-based OIRF ensemble forecasting model offers an effective alternative to conventional CTMs by promptly supplying forecasting ensemble members of $PM_{2.5}$ chemical components to the LEnKF algorithm and iteratively updating model parameters based on analysis fields derived from the LEnKF algorithm. The LEnKF algorithm effectively assimilates chemical observations into forecast fields, minimizing interference from spurious correlations by implementing localization schemes, thereby generating high-accuracy analysis fields for the OIRF model. The online coupling of the OIRF model with the LEnKF algorithm facilitates the iterative execution of ensemble forecasting, assimilation, and incremental learning at each time step. Consequently, the OIRF-LEnKF v1.0 is capable of generating high-quality forecasting and analysis fields while simultaneously undergoing self-evolution.

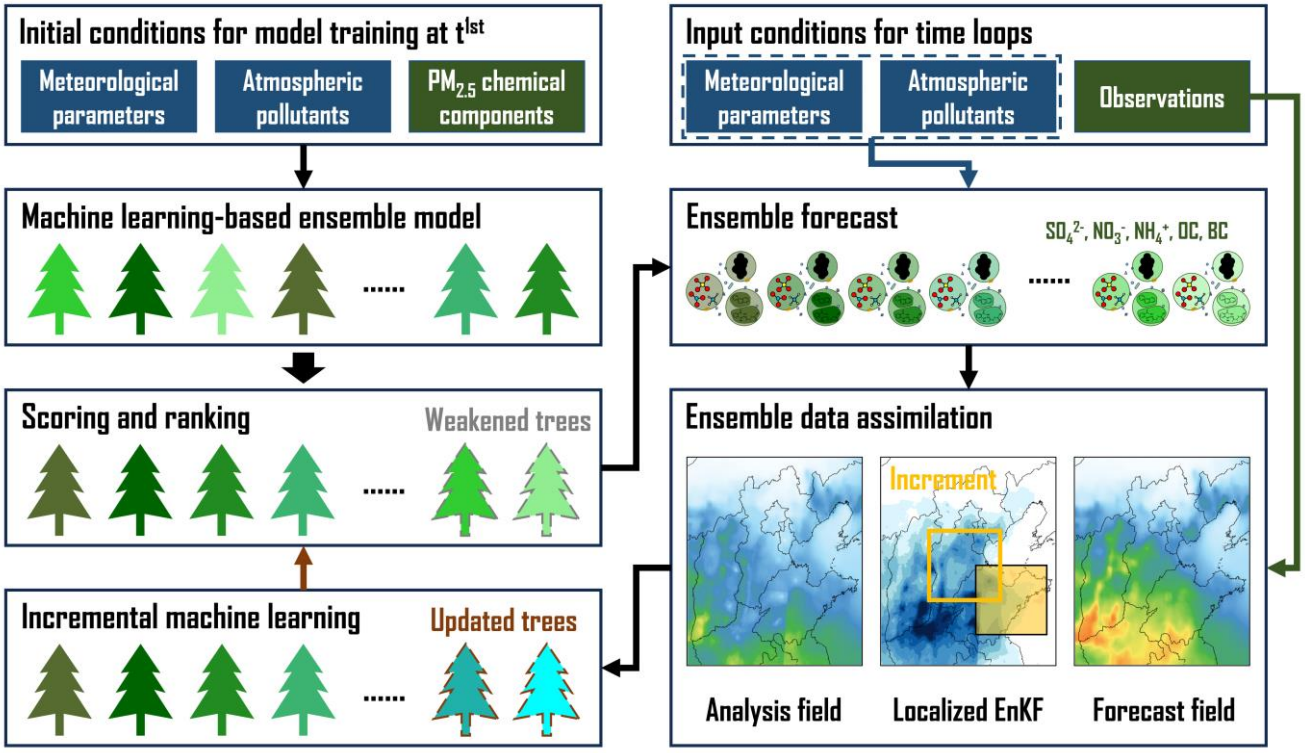


Figure 1. The framework of OIRF-LEnKF v1.0.

As shown in Fig. 1, the fundamental workflow of OIRF-LEnKF v1.0 is as follows.

Step 1. Initial training of the OIRF model. The training data at the first timestep serve as the initial conditions for constructing the OIRF model. The input features include meteorological parameters, including temperature, relative humidity, U-component wind, V-component wind, and geopotential, as well as anthropogenic atmospheric pollutants, including $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, and O_3. The output features are SO_4^{2-}, NO_3^-, NH_4^+, OC, and BC.

Step 2. Scoring the forecasting performance of ensemble decision trees in the OIRF model using mean absolute error (MAE) and screening out the decision trees with poor forecasting performance based on a predefined threshold.

Step 3. Generating a forecast ensemble of $PM_{2.5}$ chemical component concentrations at the current timestep using the OIRF model, along with the current meteorological and anthropogenic input data.

Step 4. Generating the analysis fields of $PM_{2.5}$ chemical component concentrations at the current timestep by assimilating chemical observations into forecast fields using the LEnKF algorithm.

Step 5. Incremental learning of the OIRF model. High-quality analysis fields at the current time step, along with current meteorological and anthropogenic input data, are employed to train a new ensemble of decision trees. The old decision trees, which exhibit poor forecasting performance, are subsequently replaced with new decision trees to enhance the forecasting accuracy and generalization ability of the OIRF model. Repeat *steps 2-5* until the end of the loop.

2.1.2 Optimized Incremental Random Forest (OIRF)

The OIRF model utilizes the Random Forest (RF) algorithm to establish a mapping relationship between anthropogenic atmospheric pollutants ($PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, and O_3), meteorological conditions (temperature, relative humidity, U-component wind, V-component wind, and geopotential), and the five $PM_{2.5}$ chemical components (SO_4^{2-}, NO_3^-, $NH4^+$, OC, and BC). The RF model consists of N decision trees (DTs), each using an independently and identically distributed random vector (θ_n) to facilitate feature random selection and sample bootstrapping. This approach enhances the diversity among DTs while maintaining the forecasting capability of each DT (Breiman, 2001). Unlike conventional ensemble forecasts that rely on multiple CTMs, RF can swiftly generate an ensemble of forecasting members required for DA from multiple DTs without requiring external ensemble perturbation. The forecast fields of the RF model represent the average of all DT outputs (Eq. (1)).

$$f^{RF}(x) = \frac{1}{\mathrm{N}}\sum_{\mathrm{n}=1}^{\mathrm{N}} f^{DT}(x, \theta_{\mathrm{n}}) , \quad (1)$$

Where x represents the input features, including anthropogenic atmospheric pollutants and meteorological conditions. $f^{RF}(x)$ denotes the forecast field of $PM_{2.5}$ chemical component concentrations. N is the total number of DTs. $f^{DT}(x, \theta_{\mathrm{n}})$ denotes the forecasting output of the n^{th} DT and θ_{n} is an independently and identically distributed random vector that facilitates feature random selection and sample bootstrapping.

The OIRF model incorporates a self-evolving mechanism into the RF model, enabling it to conduct incremental learning from newly available training data. In the self-evolving mechanism, the OIRF model scores the forecasting performance of each DT based on the mean absolute error (MAE), as shown in Eq. (2). The MAE is quantified by the forecast fields and high-accuracy analysis fields at the same time step.

$$f_{\mathrm{n}}^{score} = \frac{1}{\mathrm{K}}\sum_{\mathrm{i}=1}^{\mathrm{K}}|y_{\mathrm{i}} - f^{DT}(x_{\mathrm{i}}, \theta_{\mathrm{n}})|\,, \mathrm{n} = 1, 2, \ldots, \mathrm{N}\,, \quad (2)$$

Here, f_{n}^{score} is the MAE value of the nth DT. K is the total number of grids of $PM_{2.5}$ chemical component concentrations. y_{i} is the analysis value of concentrations at the nth grid point after DA. $f^{DT}(x_{\mathrm{i}}, \theta_{\mathrm{n}})$ denotes the forecasting value of the nth DT at the nth grid point.

The self-evolving mechanism introduces a threshold (τ_{p}) to screen out the DTs with poor forecasting performance. The threshold is defined as the pth percentile value of f_{n}^{score}. As shown in Eq. (3), the old DTs with scores not higher than τ_{p} are retained, while the old DTs with scores higher than τ_{p} will be replaced by new DTs obtained from the incremental learning process.

$$f_{new}^{DT} = \begin{cases} f^{DT}(y_{init}|x, \theta_{\mathrm{a}}), f_{\mathrm{a}}^{score} \le \tau_{\mathrm{p}}, \mathrm{a} = 1, 2, \ldots, \mathrm{N_p} \\ f^{DT}(y_{ana}|x, \theta_{\mathrm{b}}), f_{\mathrm{b}}^{score} > \tau_{\mathrm{p}}, \mathrm{b} = \mathrm{N_p} + 1, \mathrm{N_p} + 2, \ldots, \mathrm{N} \end{cases}, \quad (3)$$

Here, f_{new}^{DT} represents the final forecasting output of the updated DTs following incremental learning. $f^{DT}(y_{init}|x, \theta_{\mathrm{a}})$ denotes the forecasting output of the retained old DTs while $f^{DT}(y_{ana}|x, \theta_{\mathrm{b}})$ refers to the forecasting output of the new DTs. τ_{p} indicates the pth percentile value of f_{n}^{score}, and $\mathrm{N_p}$ signifies the number of retained old DTs that achieve a score not exceeding τ_{p}. The p is set at 80 to prevent excessive updating of DTs, which may introduce instability into ensemble forecasts of the OIRF model.

The final forecast field ($f^{OIRF}(x)$) of the OIRF model is derived from Eq. (4) by averaging the forecasting outputs of the updated DTs.

$$f^{OIRF}(x) = \frac{1}{\mathrm{N}}\sum_{\mathrm{n}=1}^{\mathrm{N}} f_{new}^{DT}(x, \theta_{\mathrm{n}})\,, \quad (4)$$

The self-evolving mechanism enhances the capacity of the OIRF model to incorporate newly available training data, thereby improving its generalization ability in forecasting $PM_{2.5}$ chemical component concentrations. Concurrently, the elimination of old DTs with poor forecasting performance further increases the accuracy of the forecast fields.

The hyperparameters in the OIRF model, such as the minimum number of leaf node observations, the maximal number of decision splits, and the number of predictors to select at random for each split, control the model structure and randomness level (Probst et al., 2019). The OIRF model integrates the RF model with the Bayesian optimization algorithm to ensure the statistical optimization of the hyperparameters. The Bayesian optimization algorithm incorporates hyperparameters as decision variables within the objective function, thereby abstracting the optimization problem as a solution problem of the objective function (Wu et al., 2019). This algorithm is capable of identifying the global optimal solution using fewer iterations, thereby reducing the computational costs associated with evaluating the loss function and enhancing the performance of the ML model (Shahriari et al., 2016). A probabilistic surrogate model and an acquisition function are two essential components of the Bayesian optimization algorithm. The former is employed to approximate the complex objective function, thereby minimizing computational costs. The latter is used to identify potential optimal decision variables and update the surrogate model during iterative optimization. In this study, the surrogate model and acquisition function are specifically implemented using a non-parametric Gaussian process regression model (Rasmussen, 2003, February) and the Expected Improvement per Second Plus (Elps+) function (Gelbart et al., 2014). The detailed implementation of the Bayesian optimization algorithm in machine learning models is described in our previous work (Li et al., 2025).

2.1.3 Localized Ensemble Kalman Filter (LEnKF)

LEnKF is an Ensemble Kalman Filter (EnKF) algorithm with localization schemes that mitigate filter divergence induced by sampling errors of the estimated error covariance matrix (Nerger et al., 2012), thereby generating high-precision analysis fields of $PM_{2.5}$ chemical component concentrations. The EnKF is an extension of the Kalman filter, specifically designed for atmospheric and oceanic DA with nonlinear and high-dimensional model state spaces (Houtekamer and Zhang, 2016). The EnKF utilizes the Monte Carlo method to estimate a flow-dependent background error covariance matrix from an ensemble of model states at each time step. This algorithm mitigates the high computational costs associated with the explicit operations of high-dimensional matrices (Evensen, 1994; Evensen, 2003). In this study, the OIRF model replaced the conventional CTMs to provide an ensemble of DT-based forecasting members for estimating the background error covariance (Eq. (5)). The ensemble size in DA is equal to the total number of DTs in the OIRF model.

$$\boldsymbol{P}_{\mathrm{t}}^{f}=\frac{1}{\mathrm{N}-1}\sum_{n=1}^{\mathrm{N}}\left(f_{\mathrm{t}}^{DT}(x,\theta_{\mathrm{n}})-\overline{f_{\mathrm{t}}^{DT}(x,\theta_{\mathrm{n}})}\right)\left(f_{\mathrm{t}}^{DT}(x,\theta_{\mathrm{n}})-\overline{f_{\mathrm{t}}^{DT}(x,\theta_{\mathrm{n}})}\right)^{T}, \tag{5}$$

Here, $\boldsymbol{P}_{\mathrm{t}}^{f}$ is the flow-dependent background error covariance matrix of $PM_{2.5}$ chemical component concentrations at t.

The Kalman gain matrix ($\boldsymbol{K}$) can be calculated by Eq. (6)-(8).

$$\boldsymbol{K}=\boldsymbol{P}_{\mathrm{t}}^{f}\boldsymbol{H}_{\mathrm{t}}{}^{T}\left(\boldsymbol{H}_{\mathrm{t}}\boldsymbol{P}_{\mathrm{t}}^{f}\boldsymbol{H}_{\mathrm{t}}{}^{T}+\boldsymbol{R}_{\mathrm{t}}\right)^{-1}, \tag{6}$$

$$\boldsymbol{P}_{\mathrm{t}}^{f}\boldsymbol{H}_{\mathrm{t}}{}^{T}=\frac{1}{\mathrm{N}-1}\sum_{\mathrm{n}=1}^{\mathrm{N}}\left(f_{\mathrm{t}}^{DT}(x,\theta_{\mathrm{n}})-\overline{f_{\mathrm{t}}^{DT}(x,\theta_{\mathrm{n}})}\right)\left(H\left(f_{\mathrm{t}}^{DT}(x,\theta_{\mathrm{n}})\right)-\overline{H\left(f_{\mathrm{t}}^{DT}(x,\theta_{\mathrm{n}})\right)}\right)^{T}, \tag{7}$$

$$\boldsymbol{H}_{\mathrm{t}}\boldsymbol{P}_{\mathrm{t}}^{f}\boldsymbol{H}_{\mathrm{t}}{}^{T}=\frac{1}{\mathrm{N}-1}\sum_{\mathrm{n}=1}^{\mathrm{N}}\left(H\left(f_{\mathrm{t}}^{DT}(x,\theta_{\mathrm{n}})\right)-\overline{H\left(f_{\mathrm{t}}^{DT}(x,\theta_{\mathrm{n}})\right)}\right)\left(H\left(f_{\mathrm{t}}^{DT}(x,\theta_{\mathrm{n}})\right)-\overline{H\left(f_{\mathrm{t}}^{DT}(x,\theta_{\mathrm{n}})\right)}\right)^{T}, \quad (8)$$

Here, $\boldsymbol{K}$ is the Kalman gain matrix. $\boldsymbol{H}_{\mathrm{t}}$ is the observation operator at t. $\boldsymbol{R}_{\mathrm{t}}$ is the observation error covariance matrix at t, which is a diagonal matrix. H is the linear observation operator. In this study, the observation operator solely conducts spatial mapping between the observations and the forecast fields due to consistency in the variable and temporal dimensions. The method employed for spatial mapping between observations from sparse sites and gridded forecast fields is the k-nearest neighbor search (Friedman et al., 1977).

The final analysis fields ($x_{\mathrm{n,t}}^{ana}$) can be obtained from the integration of forecast fields ($f_{\mathrm{t}}^{DT}(x,\theta_{\mathrm{n}})$) and observations ($y_{\mathrm{t}}^{o}$):

$$x_{\mathrm{n,t}}^{ana}=f_{\mathrm{t}}^{DT}(x,\theta_{\mathrm{n}})+\boldsymbol{K}\left(y_{\mathrm{t}}^{o}+y'^{o}_{\mathrm{n,t}}-H\left(f_{\mathrm{t}}^{DT}(x,\theta_{\mathrm{n}})\right)\right),\mathrm{n}=1,2,\ldots,\mathrm{N}, \quad (9)$$

Here, $x_{\mathrm{n,t}}^{ana}$ is the analysis field of the n$^{\mathrm{th}}$ ensemble member at t. y_{t}^{o} is the observation of $PM_{2.5}$ chemical components at t and $y'^{o}_{\mathrm{n,t}}$ is the observation perturbation of the n$^{\mathrm{th}}$ ensemble member at t, characterized by a normal distribution with a mean of 0 and a standard deviation equal to the observation error.

The LEnKF integrates domain localization and observation localization into the EnKF algorithm to diminish the interference of non-physical teleconnections within a high-dimensional model state space, especially for small ensemble sizes (Nerger et al., 2012). The domain localization segments the global state space into several disjoint local state spaces, each of which assimilates observations independently within a defined localization radius, thereby effectively increasing the rank of the background covariance matrix and eliminating the interference of long-distance spurious correlations (Houtekamer and Mitchell, 1998). The independence of the analysis process within the local state space facilitates parallel computation (Janjić et al., 2011). However, this may result in discontinuities at the boundaries of adjacent local state spaces. To address this challenge, domain localization in our system conducts assimilation within a specific localization radius for each analysis grid point (Fig. 2). The overlap of observations across analysis grid points smooths the boundaries of adjacent local state spaces. However, grid-by-grid assimilation at a fine spatial resolution incurs high computational costs. To mitigate this issue, OIRF-LEnKF v1.0 incorporates a second-level parallel computational framework that facilitates the simultaneous assimilation of various chemical species and multiple analysis grid points (Fig. 2). Computational tasks for different chemical species are allocated to independent computational nodes to prevent interference of spurious correlations among chemical species. Subsequently, the grid points of each chemical component are assigned to multiple CPUs within these independent computational nodes.

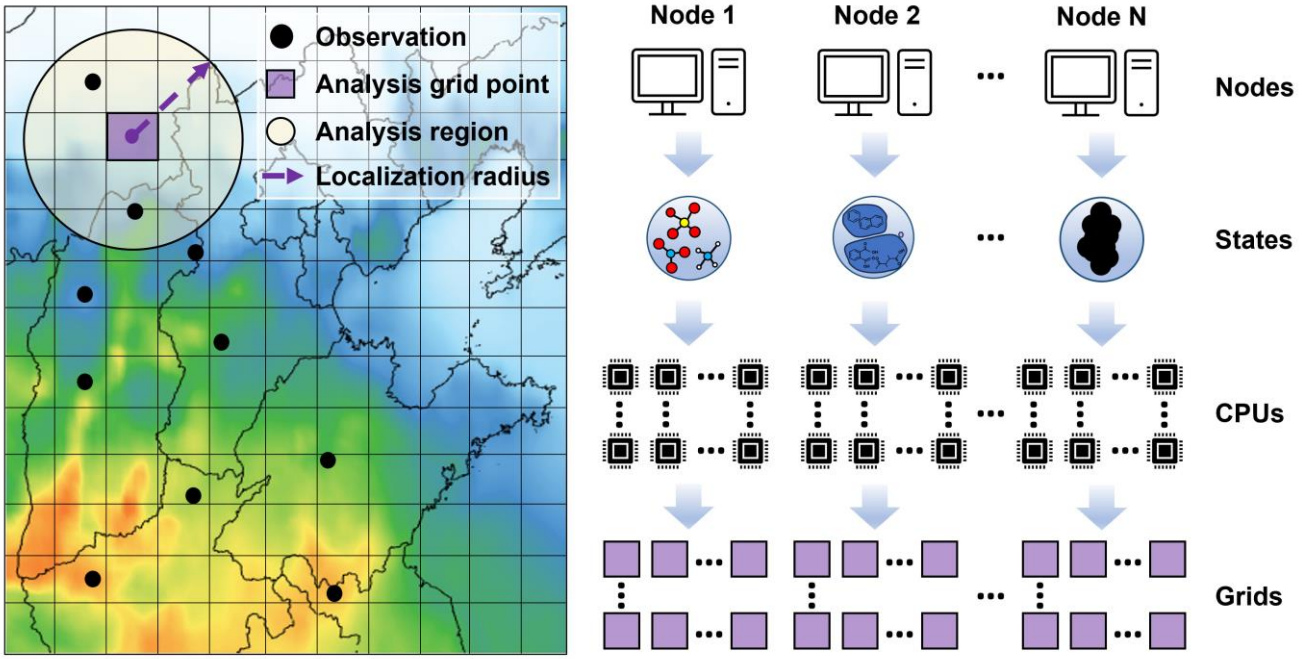


Figure 2. The scheme for domain localization and parallelization.

Observation localization is combined with domain localization to enhance the physical authenticity of observation propagation within state spaces (Nerger et al., 2012). This scheme conducts observation localization by applying the Schur product between the observation error covariance matrix ($\boldsymbol{R}_t$) and a distance-based weight matrix ($\boldsymbol{W}$) as shown in Eq. (10).

$$\boldsymbol{K}^L = \boldsymbol{P}_t^f \boldsymbol{H}_t{}^T \left(\boldsymbol{H}_t \boldsymbol{P}_t^f \boldsymbol{H}_t{}^T + \boldsymbol{W} \cdot \boldsymbol{R}_t\right)^{-1}, \quad (10)$$

Here, $\boldsymbol{K}^L$ is the Kalman gain matrix applied observation localization, and $\boldsymbol{W}$ is a distance-based weight matrix, which is diagonal.

The distance-based weight matrix ($\boldsymbol{W}$) is obtained using a Gaussian function:

$$\boldsymbol{W} = diag\left(exp\left(\frac{-d(i,j)^2}{2L^2}\right)\right), \quad (11)$$

Here, $d(i,j)$ is the Euclidean distance between grid point i and observation point j. L is the decorrelation length.

2.1.4 Configurations

Table 1 presents the fundamental configuration parameters in OIRF-LEnKF v1.0. The state variables consist of five $PM_{2.5}$ key chemical components (SO_4^{2-}, NO_3^-, NH_4^+, OC and BC). The modeling domain encompasses North China, with a spatial

range of 32.38-44.90 °N and 108.07-127.01 °E. The spatial and temporal resolutions are established at 5 km × 5 km and 1 hour, respectively. The data of the input feature utilized for training the OIRF forecasting model are outlined in Sec. 2.2.1, including U-component wind, V-component wind, temperature, relative humidity, geopotential, and the mass concentrations of $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, and O_3. The ensemble sizes employed in the assimilation experiments are 2, 5, 10, 15, 20, 30, 40, 50, 100, and 200. The update frequencies for incremental learning in the experiments include 0 (no update), 18-hour intervals, 12-hour intervals, 6-hour intervals, and 1-hour intervals. The experimental design is detailed in Sec. 2.3. Hyperparameters in the OIRF model, such as the minimum number of leaf node observations, the maximum number of decision splits, and the number of predictors to select at random for each split, are tuned using Bayesian optimization over 30 iterations. The training data are re-partitioned at each optimization iteration to enhance the robustness of the OIRF model. Regarding the DA-related parameters, the localization radius and decorrelation length are set to 200 km and 80 km, respectively, based on the spatial range and resolution requirements. The assimilation frequency matches the temporal resolution of 1 hour.

Table 1. Fundamental configuration parameters in OIRF-LEnKF v1.0.

Category	Parameter	Setting
Ensemble forecast	State variable	SO_4^{2-}, NO_3^-, NH_4^+, OC and BC
	Model domain	North China (32.38°N -44.90°N, 108.07°E-127.01°E)
	Spatial resolution	5 km×5 km
	Temporal resolution	1 h
	Meteorological input feature	U-component wind, V-component wind, temperature, relative humidity and geopotential
	Anthropogenic input feature	$PM_{2.5}$, PM_{10}, SO_2, NO_2, CO and O_3
	Ensemble size	2, 5, 10, 15, 20, 30, 40, 50, 100, 200
	Update frequency	0, 18-h interval, 12-h interval, 6-h interval, 1-h interval
	Hyperparameter for tuning	Minimum number of leaf node observations, maximal number of decision splits, and number of predictors to select at random for each split
	Optimization iteration	30
	Data partition	Re-partition at every iteration
Data assimilation	Algorithm	LEnKF
	Localization radius	200 km
	Decorrelation length	80 km
	Assimilation frequency	1 h

2.2 Data

2.2.1 Features

The input features used in the OIRF model training include six anthropogenic air pollutants and five meteorological parameters (Table 1). The hourly gridded data of anthropogenic air pollutants were obtained from Chinese Air Quality ReAnalysis (CAQRA, https://doi.org/10.11922/sciencedb.00053, last access: 17 April 2025). CAQRA is generated by assimilating surface observations of hourly concentrations of conventional air pollutants into the Nested Air Quality Prediction Modeling System (NAQPMS), with a spatial resolution of 15 km × 15 km and a 5-fold cross-validation R^2 of 0.52-0.81 (Kong et al., 2021). The hourly gridded data of meteorological parameters were obtained from the 5th Generation ECMWF ReAnalysis (ERA5, https://cds.climate.copernicus.eu/datasets, last access: 17 April 2025) with a horizontal resolution of 0.25° × 0.25°. The output features include five $PM_{2.5}$ chemical components (NH_4^+, NO_3^-, SO_4^{2-}, OC and BC). The hourly gridded data of these components were obtained from the $PM_{2.5}$ chemical composition dataset (CAQRA-aerosol, https://doi.org/10.12423/capdb_PKU.2023.DA, last access: 17 April 2025). CAQRA-aerosol is developed based on a CTM-based simulation method with an improved inorganic aerosol module and a constrained emission inventory, with a spatial resolution of 15 km × 15 km and a mean bias of less than 1.1 μg m^{-3} (Kong et al., 2025). Due to consideration of the distribution of available ground-based observational sites for $PM_{2.5}$ chemical components, the gridded data containing various features in China have been transformed into a new grid with a spatial resolution of 5 km × 5 km in North China, utilizing a triangulation-based linear interpolation method (Amidror, 2002).

2.2.2 Observations

Observations of hourly mass concentrations of five $PM_{2.5}$ chemical components (NH_4^+, NO_3^-, SO_4^{2-}, OC, and BC) were collected over a two-month period (February to March 2022) from 33 ground-based sites in North China and its surrounding areas. Of these 33 sites, 24 sites (designated as DA sites) were employed for DA and internal validation, while the remaining 9 sites (defined as VE sites) were used for independent verification to evaluate the influence of DA sites on neighboring areas. The description of site distribution and the division method of DA sites and VE sites were detailed in our previous work (Li et al., 2024).

2.2.3 Reanalysis dataset for comparison

The multi-source reanalysis datasets of $PM_{2.5}$ chemical components were collected to assess the relative quality of the reanalysis dataset generated by OIRF-LEnKF v1.0, including the CAQRA-aerosol, the Tracking Air Pollution in China (TAP, http://tapdata.org.cn/, last access: 2 June 2025), the Copernicus Atmosphere Monitoring Service ReAnalysis (CAMSRA, https://ads.atmosphere.copernicus.eu/, last access: 2 June 2025), the Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2, https://disc.gsfc.nasa.gov/datasets?project=MERRA-2, last access: 2 June 2025) and the reanalysis dataset generated by NAQPMS-PDAF v2.0 (NP2, https://doi.org/10.5281/zenodo.10886914, last

access: 2 June 2025). The High-resolution and High-quality Air Pollutants dataset for China (CHAP, https://doi.org/10.5281/zenodo.10011898, last access: 2 June 2025) was not considered in this study because it did not cover the observation period. The properties of the multi-source reanalysis datasets are presented in Table 2.

Table 2. Properties of the multi-source reanalysis datasets for $PM_{2.5}$ chemical components.

Dataset	Chemical species	Period	Temporal resolution	Vertical resolution	Spatial coverage	Spatial resolution
CAQRA-aerosol	SO_4^{2-}, NH_4^+, NO_3^-, OC, BC	2013-2022	1-hourly	Surface level	China	15 km×15 km
TAP	SO_4^{2-}, NH_4^+, NO_3^-, OM, BC	2000-present	Daily	Surface level	China	10 km×10 km
NP2	SO_4^{2-}, NH_4^+, NO_3^-, OC, BC	Feb. 2022	1-hourly	Surface level	North China	5 km×5 km
CAMSRA	NO_3^-, NH_4^+	2003-2024	3-hourly	Pressure level	Global	0.75°×0.75°
MERRA-2	SO_4^{2-}, OM, BC	1980-present	1-hourly	Surface level	Global	0.5°×0.625°

2.3 Experimental setting

We designed four experiments to evaluate the performance of OIRF-LEnKF v1.0 on forecast and analysis fields of the concentrations of SO_4^{2-}, NO_3^-, NH_4^+, OC, and BC. In the first experiment, we conducted model training, forecasting, and assimilation at the first time step using 10 distinct ensemble sizes (2, 5, 10, 15, 20, 30, 40, 50, 100, and 200) to assess the dependence of computational efficiency on ensemble size. In the second experiment, we performed 24-timestep forecasting and assimilation across 30 different scenarios, which comprised all possible combinations of 6 ensemble sizes (20, 30, 40, 50, 100, and 200) and 5 varied update frequencies for incremental learning (no update, 18-h interval, 12-h interval, 6-h interval, and 1-h interval). This design aimed to evaluate the sensitivity of forecasting and assimilation performance to ensemble size and update frequency. In the third experiment, we conducted a 2-month forecasting assimilation using ground-level observations at 24 DA sites to comprehensively assess the capabilities of OIRF-LEnKF v1.0 in interpreting the spatiotemporal distribution of $PM_{2.5}$ chemical component concentrations. In the fourth experiment, we simultaneously assimilated all ground-level observations at 33 sites to generate a 1-month reanalysis dataset of $PM_{2.5}$ chemical component concentrations in North China and compared it with multiple reanalysis datasets. The observation errors in the four experiments were set at 0.5 μg m^{-3} (NH_4^+), 0.5 μg m^{-3} (NO_3^-), 1.0 μg m^{-3} (SO_4^{2-}), 3.0 μg m^{-3} (OC), and 0.5 μg m^{-3} (BC), with the assumption that the observation errors were spatially isotropic in state space to reduce computational complexity.

3 Results and discussion

3.1 Computational efficiency

As shown in Fig. 3, we evaluate the computational efficiencies of hyperparameter tuning, forecasting and assimilation. Previous studies have indicated that the Bayesian optimization algorithm is both efficient and stable for hyperparameter tuning in various ML models (Lai, 2024). In this section, we validate its stability within the OIRF model and computational costs. Figure 3a demonstrates that both the estimated and observed minimum objective values initially decrease rapidly and subsequently converge within 10 iterations across all ensemble sizes, indicating the convergence stability and high efficiency of the OIRF model. In addition, the consistency in both the magnitude and variation between the estimated and observed minimum objective values suggests that the surrogate model employed in Bayesian optimization exhibits a high fitting accuracy for the objective function. Although the time consumed during each iteration increases positively with ensemble size, the number of optimal hyperparameter searches remains relatively insensitive to ensemble size. As illustrated in Figure 3b, the minimum value of the total observed objectives decreases significantly as the ensemble size increases, ranging from 2 to 20, indicating that a larger ensemble size enhances the optimization accuracy of the OIRF model. Notably, when the ensemble size exceeds 20, the rate of improvement in optimization accuracy diminishes. The total time consumed by the optimization process increases gradually with ensemble sizes ranging from 2 to 50 but rises sharply beyond an ensemble size of 50. Therefore, an ensemble size of 50 is determined to be optimal for the OIRF model, effectively balancing the optimization accuracy and efficiency.

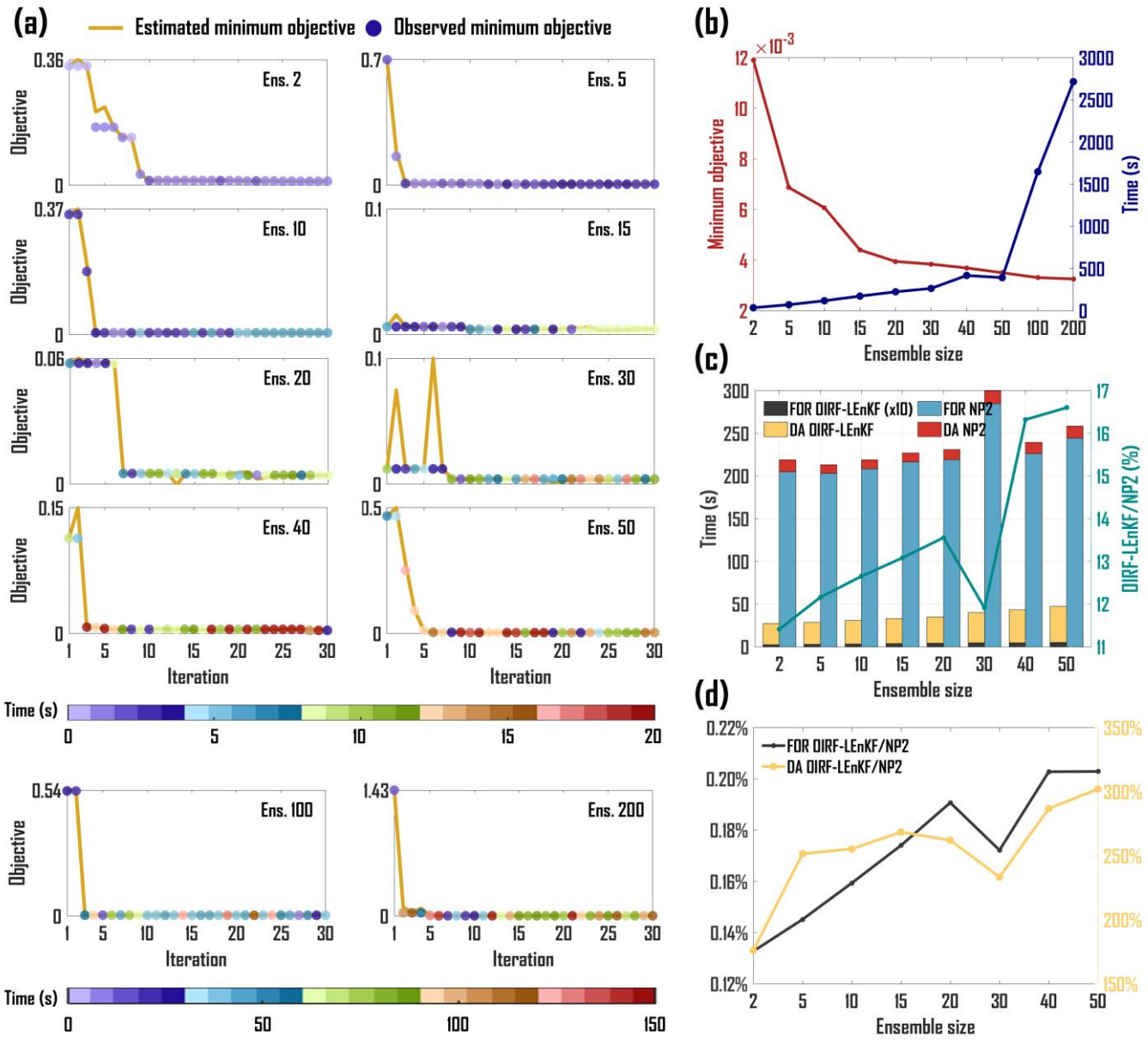


Figure 3. Computational efficiency of OIRF-LEnKF v1.0. (a) Variation in the minimum objective value throughout the Bayesian optimization process and time consumed by each iteration, (b) minimum value of total observed minimum objectives and total time consumed during Bayesian optimization process for different ensemble sizes, (c) time consumed by model forecasting and data assimilation at each timestep for OIRF-LEnKF and NAQPMS-PDAF v2.0 (NP2), and the ratio of total time consumed between OIRF-LEnKF and NP2, (d) the ratio of time consumed by model forecasting and data assimilation between OIRF-LEnKF and NP2. FOR represents the forecast phase, and DA represents the data assimilation phase. The elapsed time of the OIRF-LEnKF forecast process in Figure 3c has been magnified by a factor of 10 for better clarity.

The computational costs of OIRF-LEnKF v1.0 in forecasting and assimilation processes were compared with those of a CTM-based DA system (NP2). To ensure comparability of computational expenses between OIRF-LEnKF v1.0 and NP2, the number of CPUs allocated for each grid calculation was intentionally set closer, at 35 and 50, respectively. As illustrated in Fig. 3c, the total time consumed by forecasting and assimilation for OIRF-LEnKF v1.0 amounts to only 11.41% to 16.60% of that for NP2, especially during the forecasting process, which accounts for merely 0.13% to 0.20% (Fig. 3d). The marked improvement in forecasting efficiency by OIRF-LEnKF v1.0 is comparable to the deep neural network-based forecasting model (Adie et al., 2024). This enhancement is primarily attributed to the fact that ML-based forecasting does not necessitate a profound understanding of the complex physicochemical mechanisms of the atmosphere (Fang et al., 2022), whereas CTM-based forecasting involves intricate computations of a large number of chemical species and reaction processes (Zaveri and Peters, 1999; Stockwell et al., 1990). The computational efficiency of OIRF-LEnKF v1.0 during the DA stage is slightly lower than that of NP2, as its time consumed is 1.76 to 3.02 times greater than that of NP2 (Fig. 3d), primarily due to minor differences in the DA algorithm and the number of CPUs allocated.

As the ensemble size increases from 2 to 50, the total time consumed for OIRF-LEnKF v1.0 and NP2 increases by 17.91 s and 39.53 s, respectively. Specifically, the time consumed by forecasting increases by 0.22 s and 39.53 s, respectively, while the time consumed by assimilation increases by 17.69 s and 0 s, respectively. Although the time consumed by assimilation for OIRF-LEnKF v1.0 is sensitive to ensemble size, the total time consumed remains relatively low (less than 50 s) at an ensemble size of 50. Given that the ensemble spread typically correlates positively with ensemble size (Lei and Whitaker, 2017), configuring an ensemble size of 50 in OIRF-LEnKF v1.0 offers an optimal balance among optimization accuracy, optimization efficiency, time consumed by forecasting and assimilation, and ensemble spread.

3.2 Sensitivity to parameterization scheme

The ensemble size and update frequency for incremental learning are critical parameters that influence the forecasting and analysis capabilities of OIRF-LEnKF v1.0. Specifically, the ensemble size affects the estimation of the background error covariance matrix (Valler et al., 2019), which determines the observation propagation at the analysis step and the uncertainty range of the ensemble forecast at the forecast step. The update frequency for incremental learning drives the adaptability of the ML-based forecasting model to non-stationary data distributions (Shaheen et al., 2022), thereby influencing the generalization ability at the forecast step and indirectly affecting the background error information at the analysis step.

During the ML forecast process, the statistical indicators that compare the forecast fields and observations for OIRF-LEnKF v1.0 exhibit a pronounced sensitivity to update frequency but are less sensitive to ensemble size. With a fixed ensemble size, the correlation coefficient (CORR) increases as the update frequency rises (Fig. 4a). At the same time, the root mean square error (RMSE) decreases significantly with a higher update frequency (Fig. 4b). Specifically, the CORR rises by 2.28 % to 11.75 %, and the RMSE decreases by 32.94 % to 40.98 % when comparing a 1-hour update frequency to the scenario

without incremental learning, which indicates that high-frequency incremental learning effectively enhances the adaptability of the statically trained ML model to the non-stationary data distributions, enabling it to demonstrate improved generalization capabilities and higher forecast accuracy in rapidly changing chemical component forecasts. Notably, an increase in ensemble size can amplify the effect of incremental learning on forecast errors. Specifically, the reduction in RMSE at an ensemble size of 100 is approximately 8% greater than at an ensemble size of 20 when comparing a 1-hour update frequency to a scenario without incremental learning, which is attributed to the fact that as the ensemble size increases, the probability density distribution becomes more accurate, leading to improved ensemble forecast skill (Chen, 2024).

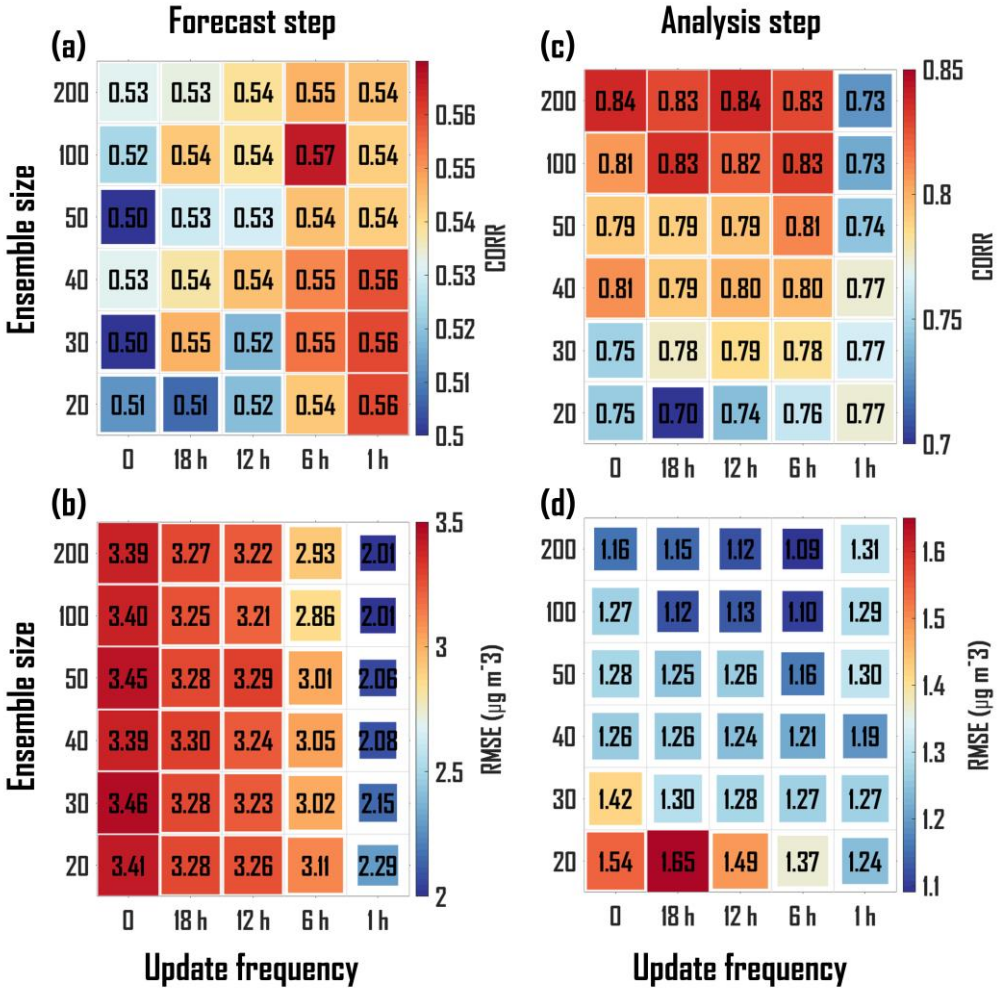


Figure 4. (a) Pearson correlation coefficient (CORR) for sensitivity test with six ensemble sizes (20, 30, 40, 50, 100, 200) and five update frequencies (no update, 18-hour interval, 12-hour interval, 6-hour interval and 1-hour interval) at the forecast step. (b) Same as (a) but for root mean square error (RMSE) at the forecast step. (c) Same as (a) but for CORR at the analysis step. (d) Same as (a) but for RMSE at the analysis step.

During the DA analysis phase, the statistical indicators that compare the analysis fields and observations for OIRF-LEnKF v1.0 are found to be significantly dependent on the ensemble size rather than the update frequency. With a fixed update frequency, excluding the 1-hour update frequency, the CORR increases considerably with a larger ensemble size (Fig. 4c). At the same time, the RMSE decreased markedly as the ensemble size increases (Fig. 4d). Specifically, the CORR increased by 8.94 % to 19.04 %, and the RMSE decreased by 20.15 % to 30.48 % when comparing an ensemble size of 200 to that of 20. This improvement is attributed to the enhanced accuracy of estimating the background error covariance matrix, resulting from a larger ensemble size, which enables the effective propagation of observations within the model state space. (Valler et al., 2019). However, the 1-hour update frequency diminishes the dependence of the analysis fields on the ensemble size. This interference may result from high-frequency incremental learning, which causes the new DTs in the OIRF model to diverge from the existing DTs, leading to a deviation in the background error covariance structure from the true state. Consequently, although the 1-hour update frequency can significantly enhance the forecasting performance, we configured an ensemble size of 50 with a 6-hour update frequency in OIRF-LEnKF v1.0 to balance computational efficiency, ML forecasting accuracy, and DA analysis performance.

3.3 Evaluation of DA results

This section assesses the performance of the background control field without DA and incremental learning (BAC), the forecast field with incremental learning (FOR) and the analysis field with DA (ANA) in interpreting the spatiotemporal distribution of $PM_{2.5}$ chemical components.

3.3.1 Assessment of temporal variation in chemical components

Figure 5 presents the time series of errors (observations minus OIRF-LEnKF v1.0 outputs) and statistical indicators comparing observations with BAC, FOR, and ANA across 33 ground-level sites. As illustrated in Fig. 5a1-a3, the errors of BAC for NH_4^+, NO_3^-, and SO_4^{2-} ranged from -2.30 ± 1.97 μg m^{-3} to 8.84 ± 5.04 μg m^{-3}, -7.60 ± 5.29 μg m^{-3} to 14.64 ± 17.20 μg m^{-3}, and -4.31 ± 3.81 μg m^{-3} to 9.61 ± 6.00 μg m^{-3}, respectively. The overall errors of BAC for NH_4^+, NO_3^-, and SO_4^{2-} are positive and relatively dispersed, suggesting a general underestimation of inorganic salt concentrations. Conversely, the errors of FOR concentrated to a range of -2.66 ± 4.18 μg m^{-3} to 5.18 ± 4.87 μg m^{-3} (NH_4^+), -7.17 ± 10.75 μg m^{-3} to 10.07 ± 7.48 μg m^{-3} (NO_3^-), and -1.37 ± 1.98 μg m^{-3} to 6.50 ± 4.81 μg m^{-3} (SO_4^{2-}), indicating that incremental learning enhances the ability to capture the temporal features of inorganic salt concentrations. Compared to BAC and FOR, the errors of ANA predominantly concentrated around zero over time, signifying that DA significantly enhances the capacity to interpret the temporal variation of inorganic salt concentrations. Unlike inorganic salt aerosols, the errors of BAC for OC and BC ranged from -12.18 ± 4.09 μg m^{-3} to -1.11 ± 2.78 μg m^{-3} and -5.41 ± 1.39 μg m^{-3} to -0.87 ± 0.57 μg m^{-3}, respectively, with a general overestimation of carbonaceous aerosol concentrations (Fig. 5a4, a5). The errors of FOR and ANA are relatively similar, both concentrating around zero over time due to the effects of incremental learning and DA.

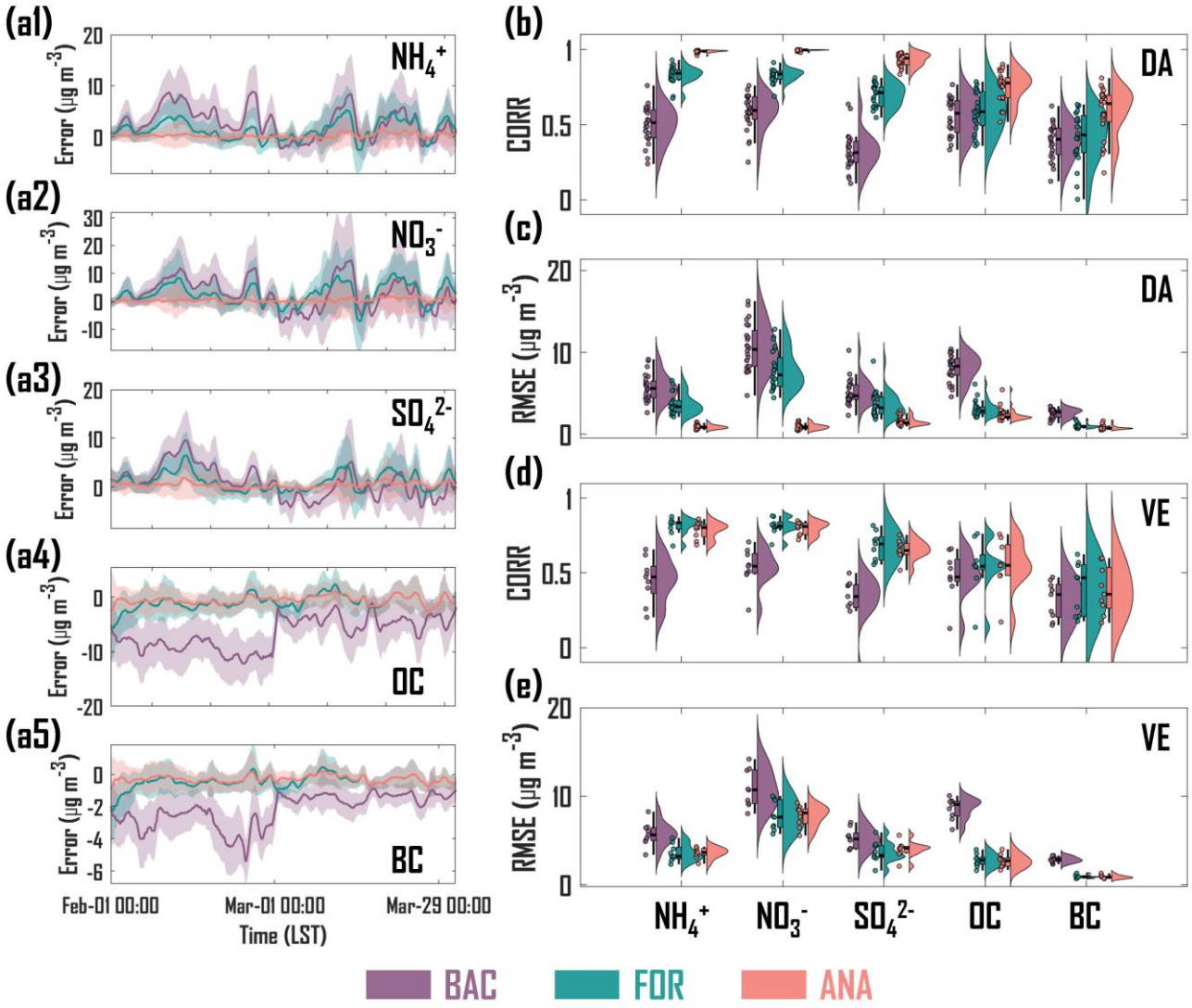


Figure 5. Smoothed variation in the error between observation and model output for (a1) NH_4^+, (a2) NO_3^-, (a3) SO_4^{2-}, (a4) OC and (a5) BC at total sites during February and March of 2022. The lines and shading areas represent the mean and standard deviation of the errors, respectively. (b) Correlation coefficient (CORR) between observation and model output for five $PM_{2.5}$ chemical components at DA sites. (c) Same as (b) but for root mean square errors (RMSE). (d) Same as (b) but for VE sites. (e) Same as (b) but for RMSE at VE sites.

Fig. 5b-e presents the CORR and RMSE for the time series of five $PM_{2.5}$ chemical components across 24 DA sites and 9 VE sites. For the DA sites, the CORR values of BAC for NH_4^+, NO_3^-, SO_4^{2-}, OC, and BC ranged from 0.24 to 0.76, 0.25 to 0.76, 0.11 to 0.64, 0.33 to 0.77, and 0.12 to 0.62, respectively (Fig. 5b). The RMSE values varied from 2.64 to 9.15 μg m^{-3}, 4.73 to

16.24 μg m^{-3}, 2.31 to 10.24 μg m^{-3}, 4.57 to 10.41 μg m^{-3}, and 1.36 to 3.42 μg m^{-3}, respectively (Fig. 5c). Following incremental learning, the CORR and RMSE values of FOR demonstrated a more concentrated data distribution than those of BAC, with average CORR (0.42 to 0.83) and RMSE (0.99 to 7.80 μg m^{-3}) values increasing by 5.61 % to 114.28 % and decreasing by 26.38 % to 61.75 %, respectively. Additionally, compared to the FOR of a CTM-based DA system, the FOR of OIRF-LEnKF v1.0 exhibited advancements of 19.14 % to 73.19 % and 33.16% to 90.10 % in CORR and RMSE, respectively (Table 3). This finding indicates that the self-evolving mechanism, characterized by incremental learning, is more effective than the optimal estimation of initial conditions in enhancing $PM_{2.5}$ chemical component forecasts, which is attributed to the fact that the enhancement in ML-based forecasting by incremental learning is global, while the CTM-based forecasting is still constrained by the uncertainties in emission inventories and physiochemical mechanisms in addition to initial conditions (Mallet and Sportisse, 2006; Luo et al., 2023). After DA, the CORR and RMSE values of ANA for NH_4^+, NO_3^-, SO_4^{2-}, OC, and BC exhibited a more concentrated data distribution than those of BAC and FOR. The average CORR (0.58 to 1.00) and RMSE (0.80 to 2.36 μg m^{-3}) values demonstrated advancements of 35.27 % to 187.15 % and 68.99 % to 91.31 %, respectively, compared to BAC, and advancements of 18.85 % to 38.73 % and 19.71 % to 88.20 %, respectively, compared to FOR.

Table 3. The correlation coefficient (CORR) and root mean square error (RMSE, μg m^{-3}) of OIRF-LEnKF v1.0 (this study) and NAQPMS-PDAF v2.0 (NP2) at DA sites and VE sites for NH4+, NO3-, SO42-, OC and BC, as well as the improvement (%) of this study relative to NP2.

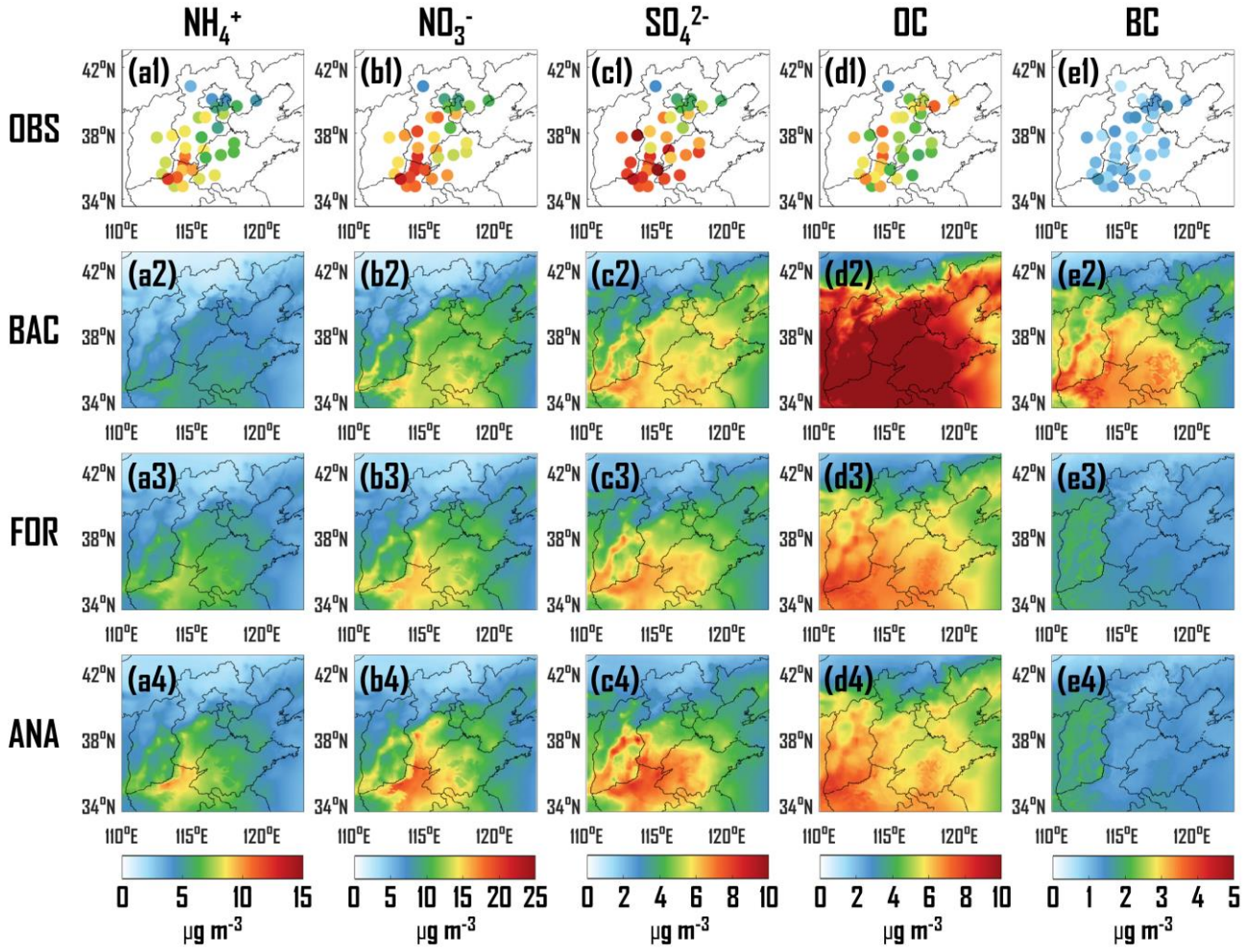
	NH_4^+		NO_3^-		SO_4^{2-}		OC		BC	
	DA	VE	DA	VE	DA	VE	DA	VE	DA	VE
	CORR									
This study	0.85	0.82	0.86	0.85	0.66	0.63	0.54	0.53	0.31	0.37
NP2	0.60	0.53	0.50	0.40	0.53	0.52	0.44	0.38	0.26	0.23
Improve (%)	41.59	53.69	73.19	110.49	23.59	21.92	23.91	41.60	19.14	64.16
	RMSE (μg m^{-3})									
This study	3.35	3.07	6.70	5.94	3.80	3.71	3.47	3.19	1.17	1.12
NP2	5.01	4.88	11.13	10.73	6.86	7.23	18.71	20.69	11.78	13.30
Improve (%)	33.16	37.10	39.77	44.62	44.59	48.73	81.48	84.58	90.10	91.55

For the VE sites without DA, the CORR values of BAC for NH_4^+, NO_3^-, SO_4^{2-}, OC, and BC ranged from 0.20 to 0.66, 0.25 to 0.71, -0.20 to 0.50, 0.13 to 0.66, and 0.15 to 0.47, respectively (Fig. 5d). The RMSE values varied from 3.39 to 8.25 μg m^{-3}, 8.04 to 14.18 μg m^{-3}, 3.94 to 7.04 μg m^{-3}, 6.23 to 10.05 μg m^{-3}, and 2.33 to 3.30 μg m^{-3}, respectively (Fig. 5e). After incremental learning, the CORR and RMSE values of FOR exhibited a more concentrated data distribution than those of BAC, with average CORR (0.39 to 0.81) and RMSE (0.93 to 7.76 μg m^{-3}) values increasing by 12.00 % to 124.69 % and decreasing by 28.37 % to 68.00 %, respectively. Furthermore, compared to the FOR of a CTM-based DA system, the FOR of OIRF-LEnKF v1.0 demonstrated advancements of 21.92 % to 110.49 % and 37.10 % to 91.55 % in CORR and RMSE,

respectively (Table 3), with greater advancements at VE sites than those at DA sites, further demonstrating the advantages of the self-evolving mechanism characterized by incremental learning for improving ML-based forecasts in a global scale. After DA, the CORR and RMSE values of ANA for NH_4^+, NO_3^-, SO_4^{2-}, OC, and BC ranged from 0.38 to 0.80 and 0.90 to 7.76 μg m^{-3}, respectively, showing a more concentrated data distribution than those of BAC and FOR. The average CORR and RMSE values increased by 14.14% to 116.65% and decreased by 23.46% to 68.75%, respectively, compared to BAC, indicating that the EnKF algorithm with localization schemes effectively propagates observations within the model state space.

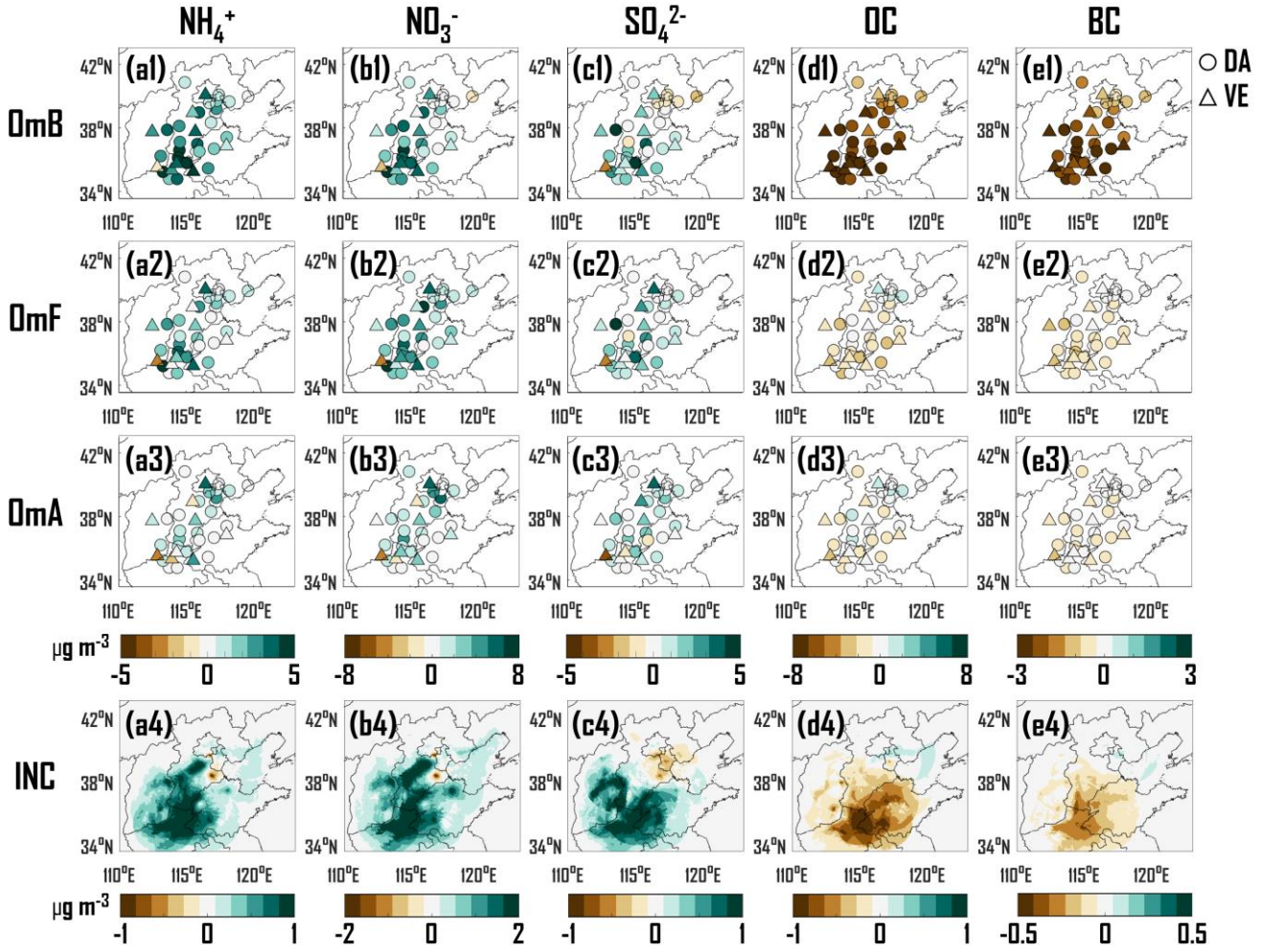
3.3.2 Assessment of spatial distribution in chemical components

Figure 6 presents the spatial distributions of observations from sparse sites (OBS), BAC, FOR and ANA for the average concentrations of NH_4^+, NO_3^-, SO_4^{2-}, OC, and BC over a two-month period from February to March 2022. The OBS of NH_4^+ reveals that the concentrations at southern sites in North China are significantly higher than those at northern sites, particularly in northern Henan Province, with a maximum concentration of 12.20 μg m^{-3} (Fig. 6a1). However, BAC fails to accurately capture the spatial patterns of NH_4^+ concentration (Fig. 6a2), exhibiting underestimations at 100 % of DA sites and 89 % of VE sites, with average underestimations of 2.71 μg m^{-3} and 3.07 μg m^{-3}, respectively (Fig. 7a1). This finding is attributed to the underestimation of the original training samples (Kong et al., 2025). Compared to BAC, the FOR mitigates the underestimation (Fig. 6a3), with 96 % of DA sites underestimating by 1.56 μg m^{-3} and 78 % of VE sites underestimating by 1.88 μg/m^3 (Fig. 7a2). After DA, ANA accurately depicts the spatial distribution of NH_4^+ concentrations (Fig. 6a4), with 92 % of DA sites underestimating by 0.74 μg m^{-3} and 44% of VE sites underestimating by 2.34 μg m^{-3}, respectively (Fig. 7a3). The increment field (INC) between ANA and FOR exhibits substantial positive increments in southern North China (Fig. 7a4), indicating that the observations from 24 DA sites were effectively propagated within the model state space, thereby addressing the underestimation of NH_4^+ concentrations in the whole domain.



510

Figure 6. Spatial distribution of observation (OBS), background control field (BAC), forecast field (FOR) and analysis field (ANA) for NH_4^+ (a1-a4), NO_3^- (b1-b4), SO_4^{2-} (c1-c4), OC (d1-d4) and BC (e1-e4).



515 **Figure 7. Spatial distribution of observation minus background control field (OmB), observation minus forecast field (OmF), observation minus analysis field (OmA) and analysis field minus forecast field (INC) for NH_4^+ (a1-a4), NO_3^- (b1-b4), SO_4^{2-} (c1-c4), OC (d1-d4) and BC (e1-e4). The circle indicates the DA sites with data assimilation, and the upward-pointing triangle indicates the VE sites without data assimilation.**

520 The observed spatial distributions of NO_3^- and SO_4^{2-} are consistent with those of NH_4^+, revealing significantly higher concentrations at southern sites in the North China region than at northern sites, particularly in the Hebei-Henan-Shandong junction areas (Fig. 6b1, c1). Although BAC can capture the spatial patterns of NO_3^- and SO_4^{2-}, it significantly underestimates their concentrations (Fig. 6b2, c2). Specifically, 63-79 % of DA sites and 89% of VE sites underestimate by 1.87-3.76 μg m^{-3} and 1.57-3.44 μg m^{-3}, respectively (Fig. 7b1, c1). Compared to BAC, FOR mitigates the underestimations
525 in the Hebei-Henan-Shandong junction areas and overestimations in the Beijing-Tianjin-Hebei eastern areas (Fig. 6b3, c3), with improvements at most DA and VE sites (Fig. 7b2, c2). After DA, ANA accurately characterizes the spatial distribution

of NO_3^- and SO_4^{2-} concentrations (Fig. 6b4, c4), with 88-100 % of DA sites and 56-67 % of VE sites merely underestimating by 0.77-1.31 μg m^{-3} and 1.85-2.73 μg m^{-3}, respectively (Fig. 7b3, c3). Furthermore, similar to the INC of NH_4^+, INCs of NO_3^- and SO_4^{2-} exhibit widespread positive increments across the North China region (Fig. 7b4, c4).

In contrast to the spatial distributions of NH_4^+, NO_3^- and SO_4^{2-}, the observed spatial distributions of OC and BC reveal that concentrations in the North China region demonstrate spatial homogeneity (Fig. 6d1, e1). However, BAC significantly overestimated the concentrations of OC and BC in the North China region (Fig. 6d2, e2, and Fig. 7d1, e1), with an average overestimation of 6.12 μg m^{-3} for OC and 1.99 μg m^{-3} for BC at all DA sites, and 6.88 μg m^{-3} for OC and 2.29 μg m^{-3} for BC at all VE sites. Following incremental learning, FOR significantly reduced the overestimations (Fig. 6d3, e3, and Fig. 7d2, e2), resulting in an average overestimation of 1.46 μg m^{-3} for OC and 0.53 μg m^{-3} for BC at 71-79 % of DA sites, and 1.56 μg m^{-3} for OC and 0.65 μg m^{-3} for BC at 89 % of VE sites. The number of sites exhibiting overestimation and the degree of overestimation are markedly lower than those of BAC. After DA, ANA further mitigates the overestimation in FOR, accurately interpreting the spatial distributions of OC and BC concentrations (Fig. 6d4, e4), with the gaps between the observations and analysis fields for both DA and VE sites approaching 0 (Fig. 7d3, e3). Assimilating the observations from 24 DA sites effectively mitigates the overestimation in the southern North China region (Fig. 7d4, e4).

3.4 Comparison with multiple reanalysis datasets

In this section, we utilized OIRF-LEnKF v1.0 to generate an hourly reanalysis dataset of $PM_{2.5}$ key chemical components (SO_4^{2-}, NO_3^-, NH_4^+, OC and BC) for the North China region in February 2022. We compared it with multiple related reanalysis datasets, including CAQRA-aerosol, TAP, Global-RA (CAMS and MERRA-2), and the dataset generated by NP2. The temporal and spatial resolutions of CAQRA-aerosol, TAP, and Global-RA on both global and national scales are lower than those of OIRF-LEnKF v1.0 and NP2 on the regional scale (Table 2). It is important to note that the spatial range and resolution of OIRF-LEnKF v1.0 are contingent upon those of the available training data. Consequently, OIRF-LEnKF v1.0 has significant potential for elucidating the spatiotemporal distribution of $PM_{2.5}$ chemical components on a global and national scale.

Figure 8 illustrates the average values of observation minus analysis (OmA) over 1 month. For NH_4^+ (Fig. 8a1-a5), the mean absolute OmA of OIRF-LEnKF v1.0 at a total of 33 sites (0.25 μg m^{-3}) is significantly lower than that of NP2 (0.81 μg m^{-3}), CAQRA (1.18 μg m^{-3}), TAP (0.92 μg m^{-3}), and Global-RA (2.92 μg m^{-3}). Furthermore, the OmA of OIRF-LEnKF v1.0 is within ±1 μg m^{-3} at 97 % of the sites, whereas NP2, CAQRA, TAP, and Global-RA had only 9-70 % of the sites within this range. Most of the sites exhibit slight underestimations in NP2 and TAP, overestimations in CAQRA, and significant underestimations in Global-RA, while the disparity between OIRF-LEnKF v1.0 and the observations is minimal. The findings for NO_3^- are comparable to those for NH_4^+ (Fig. 8b1-b5), the mean absolute OmA of OIRF-LEnKF v1.0 at a total of 33 sites (0.19 μg m^{-3}) is significantly lower than that of NP2 (0.93 μg m^{-3}), CAQRA (8.42 μg m^{-3}), TAP (2.24 μg m^{-3}), and

Global-RA (2.27 μg m^{-3}). Furthermore, the OmA of OIRF-LEnKF v1.0 is within ±2 μg m^{-3} at all sites, whereas NP2, CAQRA, TAP, and Global-RA had only 3-94 % of the sites within this range. The similar spatial patterns of OmA for NH_4^+ and NO_3^- are related to thermodynamic equilibrium (Nenes et al., 1998) and consistency between NH_4^+ and NO_3^- has also been observed in previous works (Sun, 2018; Shi et al., 2021; Wu et al., 2022).

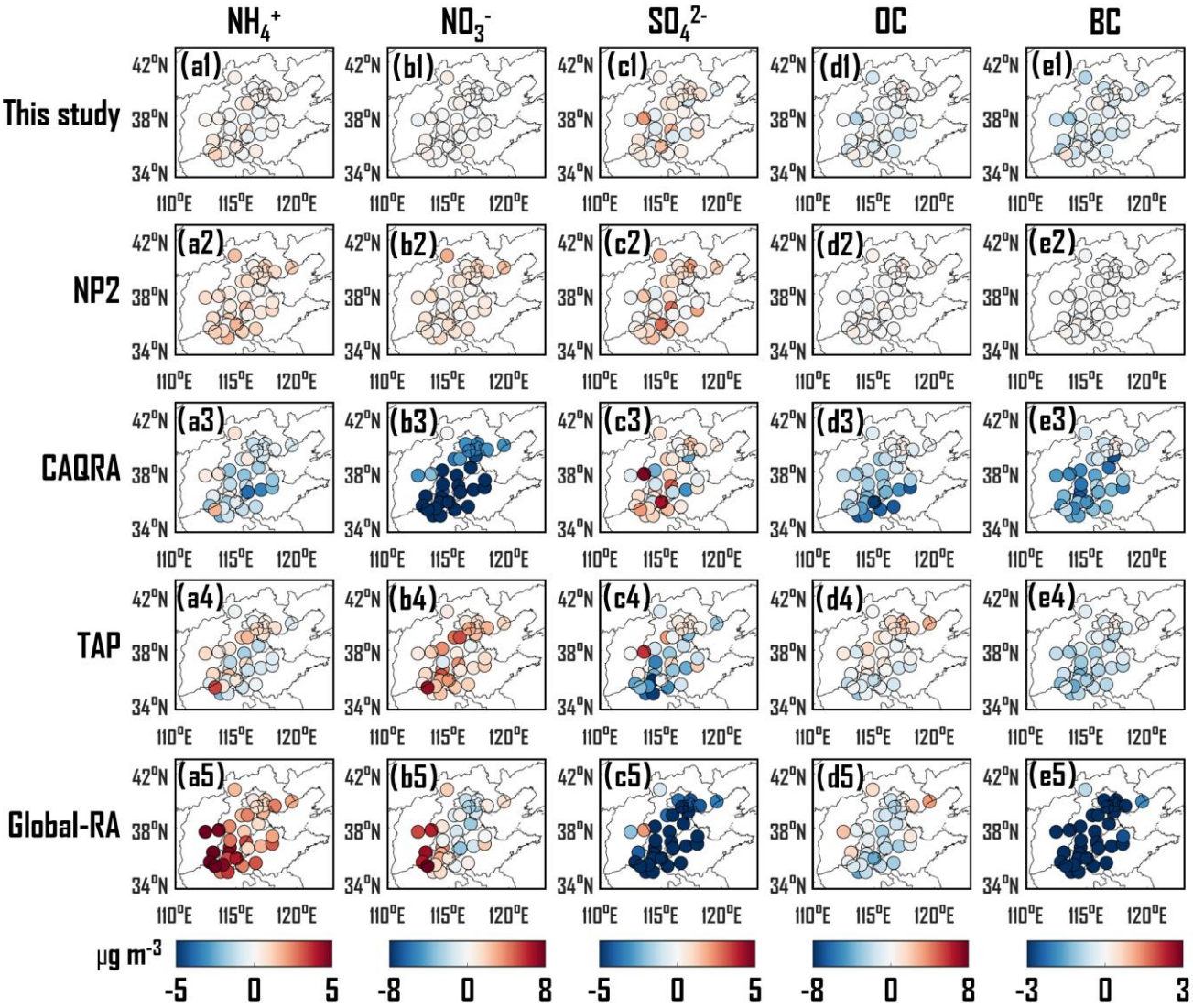


Figure 8. Difference between observations at a total of 33 sites and five reanalysis datasets for NH_4^+ (a1-a5), NO_3^- (b1-b5), SO_4^{2-} (c1-c5), OC (d1-d5) and BC (e1-e5). Global-RA is the combination of CAMSRA and MERRA-2.

For SO_4^{2-} (Fig. 8c1-c5), the average absolute OmA of OIRF-LEnKF v1.0 (0.54 μg m^{-3}) is slightly lower than that of NP2 (0.86 μg m^{-3}) but significantly lower than that of CAQRA (1.26 μg m^{-3}), TAP (1.72 μg m^{-3}), and Global-RA (7.19 μg m^{-3}). In contrast to NO_3^-, most of the sites exhibit underestimation in CAQRA, overestimation in TAP, and significant overestimation in Global-RA for SO_4^{2-}. This discrepancy between NO_3^- and SO_4^{2-} arises from the competition for the capture of NH_3. Thus, the underestimation of SO_4^{2-} is considered a factor in the overestimation of NO_3^- (Xie et al., 2022). Unlike the four CTM-based reanalysis datasets, OIRF-LEnKF v1.0 implements independent forecasting and DA processes for various chemical components, thereby reducing the constraints imposed by correlations among variables.

The OmA of OC (Fig. 8d1–d5) and BC (Fig. 8e1–e5) exhibit similar spatial patterns. Specifically, the average absolute OmA of OIRF-LEnKF v1.0 (0.66 μg m^{-3} for OC and 0.40 μg m^{-3} for BC) is slightly higher than that of NP2 (0.23 μg m^{-3} for OC and 0.03 μg m^{-3} for BC) but significantly lower than those of CAQRA (2.90 μg m^{-3} for OC and 1.32 μg m^{-3} for BC), TAP (1.04 μg m^{-3} for OC and 0.65 μg m^{-3} for BC), and Global-RA (1.62 μg m^{-3} for OC and 5.85 μg m^{-3} for BC). The significant overestimation of carbonaceous aerosols observed in CTM-based CAQRA and Global-RA is likely attributed to the hygroscopic growth schemes of carbonaceous aerosols, the poorly constrained semi-volatile species that escape from primary organic aerosols, and aging mechanisms (Soni et al., 2021; Huang et al., 2013). Overall, the reanalysis dataset generated by OIRF-LEnKF v1.0 demonstrates lower errors in the concentrations of the five $PM_{2.5}$ chemical components in the North China region compared to four CTM-based datasets.

We further compared the differences in RMSE and CORR among five reanalysis datasets. As illustrated in Fig. 9a-c, the CORR values of OIRF-LEnKF v1.0 for NH_4^+, NO_3^-, and SO_4^{2-} (mean CORR: 0.97, Fig. 9f) are significantly higher than those of other datasets (mean CORR: 0.56 to 0.89, Fig. 9f), while the RMSE values (mean RMSE: 1.12 μg m^{-3}, Fig. 9g) are significantly lower than those of other datasets (mean RMSE: 2.55-8.52 μg m^{-3}, Fig. 9g). Furthermore, the RMSE values of OIRF-LEnKF v1.0 are relatively concentrated across all sites, indicating a marked improvement in simulation of NH_4^+, NO_3^-, and SO_4^{2-} across a broad spatial range. From Fig. 9d-e, the CORR and RMSE values of OIRF-LEnKF v1.0 for carbonaceous aerosols (OC and BC) (mean CORR: 0.68, Fig. 9f; mean RMSE: 1.49 μg m^{-3}, Fig. 9g) are slightly worse than those of NP2 (mean CORR: 0.97, Fig. 9f; mean RMSE: 1.66 μg m^{-3}, Fig. 9g) and are comparable to those of TAP (mean CORR: 0.66, Fig. 9f; mean RMSE: 1.49 μg m^{-3}, Fig. 9g), while demonstrating superiority over the other datasets (mean CORR: 0.28-0.44, Fig. 9f; mean RMSE: 4.49-11.70 μg m^{-3}, Fig. 9g). Overall, OIRF-LEnKF v1.0 exhibits a notable advantage in accurately interpreting the concentrations of $PM_{2.5}$ chemical components on a regional scale. Further improvements in the performance of OIRF-LEnKF v1.0 in interpreting carbonaceous aerosols are expected by modifying the structure of the OIRF forecasting model and the frequency of incremental learning, as well as by adopting hybrid nonlinear DA algorithms.

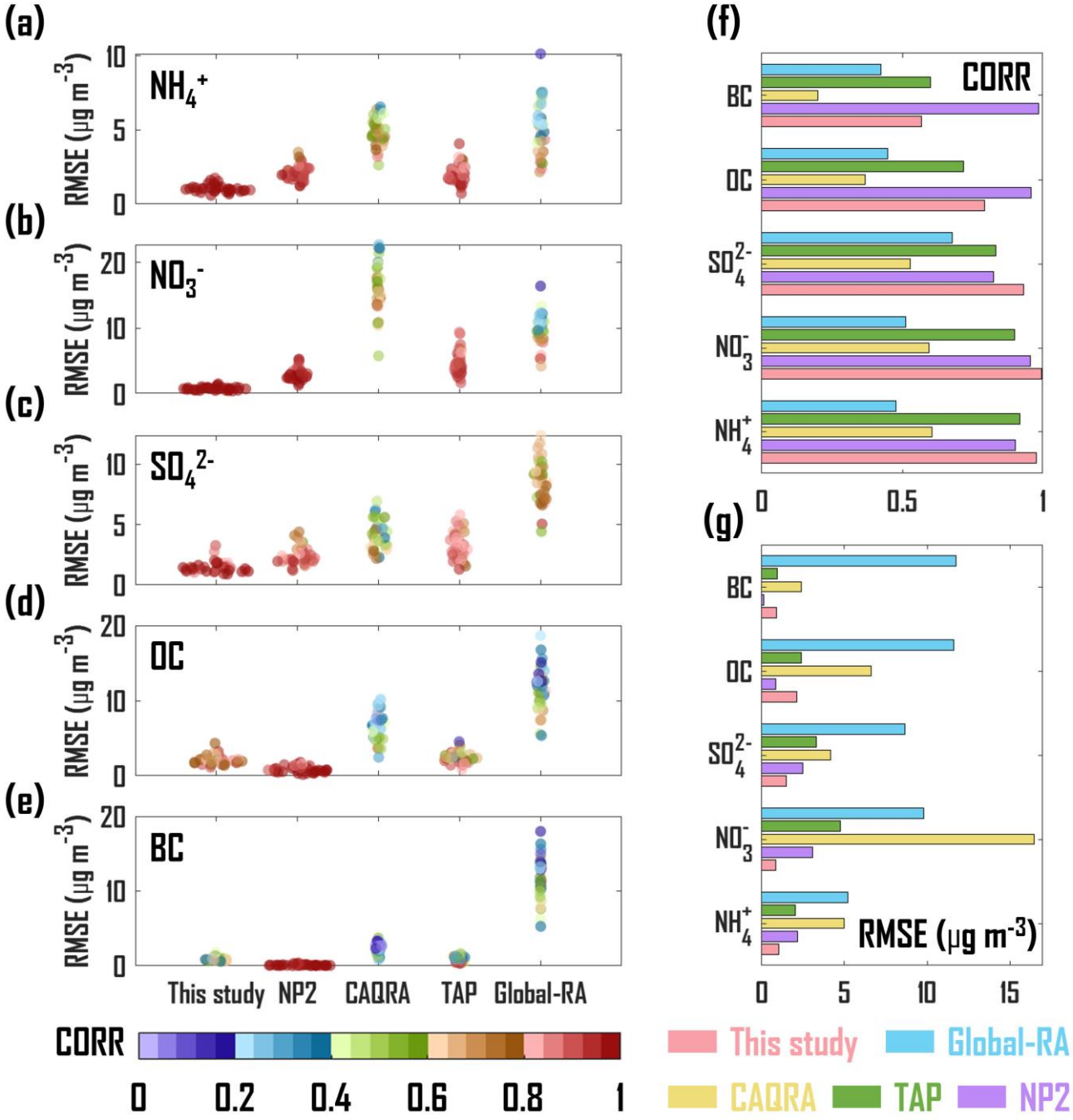


Figure 9. Pearson correlation coefficient (CORR) and root mean square error (RMSE, μg m-3) quantified by the five reanalysis datasets and observations at a total of 33 sites for NH_4^+ (a), NO_3^- (b), SO_4^{2-} (c), OC (d) and BC (e). The averages of CORR (f) and RMSE (g) across all observational sites for the five reanalysis datasets for the five PM2.5 chemical components. Global-RA is the combination of CAMSRA and MERRA-2.

4 Conclusions

In this paper, we online coupled the OIRF model with the LEnKF algorithm to develop a self-evolving DA system (OIRF-LEnKF v1.0) that mitigates the limitations of high computational costs and inadequate advancements in forecasting and analysis fields of $PM_{2.5}$ chemical components (NH_4^+, SO_4^{2-}, NO_3^-, OC and BC) in conventional CTM-based DA. The OIRF model introduces a self-evolving mechanism that enhances the generalization ability of ML by iteratively absorbing newly available training data to dynamically update the model structure. The domain localization and observation localization schemes are incorporated into the EnKF algorithm within a second-level parallel computation framework, which effectively reduces the interference of spatial and variable spurious correlations and improves computational efficiency. The findings are outlined as follows.

OIRF-LEnKF v1.0 exhibits stable convergence capability and high convergence efficiency, achieving convergence within 10 iterations across ensemble sizes ranging from 2 to 200. Computational tests reveal that the total time consumed by OIRF-LEnKF v1.0 constitutes only 11.41-16.60 % of that of CTM-based DA, particularly during the forecasting process (0.13-0.20 %), demonstrating its superior computational efficiency.

Sensitivity tests reveal that the forecast fields in OIRF-LEnKF v1.0 are more sensitive to updating frequency within the self-evolving mechanism. In contrast, the analysis fields exhibit a marked sensitivity to ensemble size. Specifically, the CORR rises by 2.28-11.75 %, and the RMSE decreases by 32.94-40.98 % when comparing a 1-hour update frequency to the scenario without incremental learning during the forecasting phase. Additionally, the CORR increases by 8.94-19.04 %, and the RMSE decreases by 20.15-30.48 % when comparing an ensemble size of 200 to that of 20 during the DA analysis phase. However, the 1-hour update frequency diminishes the dependence of the analysis fields on ensemble size. Thus, an ensemble size of 50 with a 6-hour update frequency is configured to balance computational efficiency, ML forecasting accuracy, and DA analysis performance.

A 2-month DA experiment demonstrates that the RMSE values for $PM_{2.5}$ chemical components at DA sites range from 0.99 to 7.80 $\mu g\,m^{-3}$ after incremental learning and 0.80 to 2.36 $\mu g\,m^{-3}$ after DA analysis, exhibiting reductions of 26.38-61.75 % and 68.99-91.31 %, respectively, compared to values obtained without incremental learning and DA analysis. For VE sites, the RMSE values range from 0.93 to 7.76 $\mu g\ m^{-3}$ after incremental learning and 0.90 to 7.76 $\mu g\ m^{-3}$ after DA analysis, exhibiting reductions of 28.37-68.00 % and 23.46-68.75%, respectively, relative to values obtained without incremental learning and DA analysis. Notably, the RMSE values of our system during the forecasting process show a significant reduction of 33.16-90.10 % at DA sites and 37.10-91.55 % at VE sites compared to those of CTM-based DA, highlighting the superior forecasting capability of ML-based DA. Additionally, the spatial patterns of the forecast and analysis fields for chemical components more accurately reflect those of the observations when employing incremental learning and DA.

In comparison to the datasets provided by NP2, CAQRA, TAP, CAMSRA, and MERRA-2, the dataset generated by OIRF-LEnKF v1.0 exhibits superior data quality. Notably, for NH_4^+, NO_3^- and SO_4^{2-}, the CORR values of OIRF-LEnKF v1.0 (0.97) are significantly higher than those of the aforementioned datasets (0.56-0.89). Additionally, the RMSE values of OIRF-LEnKF v1.0 (1.12 μg m^{-3}) are markedly lower than those of the four reanalysis datasets (2.55-8.52 μg m^{-3}). Future work should focus on generating reanalysis datasets that utilize configurations with larger domains and higher spatial resolutions, as well as improving data quality through the application of deep learning techniques and hybrid nonlinear DA algorithms.

Code and data availability

The source codes and related data in our work are openly accessible at https://doi.org/10.5281/zenodo.16735735 (Li et al., 2025).

Author contributions

HL developed the data assimilation system, performed numerical experiments, carried out the analysis and wrote this paper. TY provided scientific guidance and wrote this paper. LK provided help for the system code and the CAQRA reanalysis dataset. DWZ, DZ, and GT provided $PM_{2.5}$ chemical component data. HS, and ZW did overall supervision. All authors reviewed and revised this paper.

Competing interests

The contact author has declared that neither they nor their co-authors have any competing interests.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) Excellent Young Scientists Fund (No. 42422506). We thank the National Natural Science Foundation of China (No. 42275122), the technical support of the National large Scientific and Technological Infrastructure “Earth System Numerical Simulation Facility” (https://cstr.cn/31134.02. EL), and the data support of the China National Environmental Monitoring Center. Ting Yang would like to express gratitude towards the Program of the Youth Innovation Promotion Association (CAS).

Financial support

This work was supported by the National Natural Science Foundation of China (NSFC) Excellent Young Scientists Fund (No. 42422506). National Key Research and Development Program of China (No. 2023YFC3705801), the National Natural Science Foundation of China（No. 42275122).

References

Adie, J., Chin, C. S., Li, J., and See, S.: GAIA-Chem: A Framework for Global AI-Accelerated Atmospheric Chemistry Modelling, in: Proceedings of the Platform for Advanced Scientific Computing Conference, Zurich, Switzerland, 13, 1-5, https://doi.org/10.1145/3659914.3659927, 2024.

Amidror, I.: Scattered data interpolation methods for electronic imaging systems: a survey, J. Electron. Imaging, 11, https://doi.org/10.1117/1.1455013, 2002.

Arcucci, R., Zhu, J., Hu, S., and Guo, Y.-K.: Deep Data Assimilation: Integrating Deep Learning with Data Assimilation, Appl. Sci., 11, 1114, https://doi.org/10.3390/app11031114, 2021.

Brajard, J., Carrassi, A., Bocquet, M., and Bertino, L.: Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: A case study with the Lorenz 96 model, J. Comput. Sci., 44, 101171, https://doi.org/10.1016/j.jocs.2020.101171, 2020.

Breiman, L.: Random Forests, Mach. Learn., 45, 5-32, https://doi.org/10.1023/A:1010933404324, 2001.

Buizza, C., Quilodrán Casas, C., Nadler, P., Mack, J., Marrone, S., Titus, Z., Le Cornec, C., Heylen, E., Dur, T., Baca Ruiz, L., Heaney, C., Díaz Lopez, J. A., Kumar, K. S. S., and Arcucci, R.: Data Learning: Integrating Data Assimilation and Machine Learning, J. Comput. Sci., 58, 101525, https://doi.org/10.1016/j.jocs.2021.101525, 2022.

Cha, Y., Lee, J.-J., Song, C. H., Kim, S., Park, R. J., Lee, M.-I., Woo, J.-H., Choi, J.-H., Bae, K., Yu, J., Kim, E., Kim, H., Lee, S.-H., Kim, J., Chang, L.-S., Jeon, K.-h., and Song, C.-K.: Investigating uncertainties in air quality models used in GMAP/SIJAQ 2021 field campaign: General performance of different models and ensemble results, Atmos. Environ., 340, 120896, https://doi.org/10.1016/j.atmosenv.2024.120896, 2025.

Chattopadhyay, A., Nabizadeh, E., Bach, E., and Hassanzadeh, P.: Deep learning-enhanced ensemble-based data assimilation for high-dimensional nonlinear dynamical systems, J. Comput. Phys., 477, 111918, https://doi.org/10.1016/j.jcp.2023.111918, 2023.

Chen, L.: A review of the applications of ensemble forecasting in fields other than meteorology, Weather, 79, 285-290, https://doi.org/10.1002/wea.4584, 2024.

Dong, R., Leng, H., Zhao, J., Song, J., and Liang, S.: A Framework for Four-Dimensional Variational Data Assimilation Based on Machine Learning, Entropy, 24, 264, https://doi.org/10.3390/e24020264, 2022.

Dong, R., Leng, H., Zhao, C., Song, J., Zhao, J., and Cao, X.: A hybrid data assimilation system based on machine learning, Front. Earth Sci., 10, https://doi.org/10.3389/feart.2022.1012165, 2023.

Evensen, G.: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, J. Geophys. Res. Oceans, 99, https://doi.org/10.1029/94jc00572, 1994.

Evensen, G.: The Ensemble Kalman Filter: Theoretical formulation and practical implementation, Ocean Dyn., 53, 343-367, https://doi.org/10.1007/s10236-003-0036-9, 2003.

Fang, L., Jin, J., Segers, A., Lin, H. X., Pang, M., Xiao, C., Deng, T., and Liao, H.: Development of a regional feature selection-based machine learning system (RFSML v1.0) for air pollution forecasting over China, Geosci. Model Dev., 15, 7791-7807, https://doi.org/10.5194/gmd-15-7791-2022, 2022.

Farchi, A., Bocquet, M., Laloyaux, P., Bonavita, M., and Malartic, Q.: A comparison of combined data assimilation and machine learning methods for offline and online model error correction, J. Comput. Sci., 55, 101468, https://doi.org/10.1016/j.jocs.2021.101468, 2021.

Friedman, J. H., Bentley, J. L., and Finkel, R. A.: An algorithm for finding best matches in logarithmic expected time, ACM Transactions on Mathematical Software (TOMS), 3, 209-226, https://doi.org/10.1145/355744.355745, 1977.

Geer, A. J.: Learning earth system models from observations: machine learning or data assimilation?, Philosophical Transactions of the Royal Society A, 379, 20200089, https://doi.org/doi:10.1098/rsta.2020.0089, 2021.

Gelbart, M. A., Snoek, J., and Adams, R. P.: Bayesian optimization with unknown constraints, in: Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence (UAI'14), AUAI Press, Arlington, Virginia, USA, 250-259, https://dl.acm.org/doi/10.5555/3020751.3020778, 2014.

Gottwald, G. A. and Reich, S.: Supervised learning from noisy observations: Combining machine-learning techniques with data assimilation, Phys. D: Nonlinear Phenom., 423, 132911, https://doi.org/10.1016/j.physd.2021.132911, 2021.

He, X., Li, Y., Liu, S., Xu, T., Chen, F., Li, Z., Zhang, Z., Liu, R., Song, L., Xu, Z., Peng, Z., and Zheng, C.: Improving regional climate simulations based on a hybrid data assimilation and machine learning method, Hydrol. Earth Syst. Sci., 27, 1583-1606, https://doi.org/10.5194/hess-27-1583-2023, 2023.

Houtekamer, P. L. and Mitchell, H. L.: Data Assimilation Using an Ensemble Kalman Filter Technique, Mon. Weather Rev., 126, 796-811, https://doi.org/10.1175/1520-0493(1998)126<0796:DAUAEK>2.0.CO;2, 1998.

Houtekamer, P. L. and Zhang, F.: Review of the Ensemble Kalman Filter for Atmospheric Data Assimilation, Mon. Weather Rev., 144, 4489-4532, https://doi.org/10.1175/MWR-D-15-0440.1, 2016.

Howard, L. J., Subramanian, A., and Hoteit, I.: A Machine Learning Augmented Data Assimilation Method for High-Resolution Observations, J. Adv. Model Earth Sy., 16, e2023MS003774, https://doi.org/10.1029/2023MS003774, 2024.

Huang, R. J., Zhang, Y. L., Bozzetti, C., Ho, K. F., Cao, J. J., Han, Y. M., Daellenbach, K. R., Slowik, J. G., Platt, S. M., Canonaco, F., Zotter, P., Wolf, R., Pieber, S. M., Bruns, E. A., Crippa, M., Ciarelli, G., Piazzalunga, A., Schwikowski, M., Abbaszade, G., Schnelle-Kreis, J., Zimmermann, R., An, Z. S., Szidat, S., Baltensperger, U., El Haddad, I., and Prévôt, A. S. H.: High secondary aerosol contribution to particulate pollution during haze events in China, Nature, 514, 218-222, https://doi.org/10.1038/nature13774, 2014.

Huang, Y., Wu, S., Dubey, M. K., and French, N. H. F.: Impact of aging mechanism on model simulated carbonaceous

aerosols, Atmos. Chem. Phys., 13, 6329-6343, https://doi.org/10.5194/acp-13-6329-2013, 2013.

Inness, A., Ades, M., Agustí-Panareda, A., Barré, J., Benedictow, A., Blechschmidt, A. M., Dominguez, J. J., Engelen, R., Eskes, H., Flemming, J., Huijnen, V., Jones, L., Kipling, Z., Massart, S., Parrington, M., Peuch, V. H., Razinger, M., Remy, S., Schulz, M., and Suttie, M.: The CAMS reanalysis of atmospheric composition, Atmos. Chem. Phys., 19, 3515-3556, https://doi.org/10.5194/acp-19-3515-2019, 2019.

Janjić, T., Nerger, L., Albertella, A., Schröter, J., and Skachko, S.: On Domain Localization in Ensemble-Based Kalman Filter Algorithms, Mon. Weather Rev., 139, 2046-2060, https://doi.org/10.1175/2011MWR3552.1, 2011.

Jin, J., Lin, H. X., Segers, A., Xie, Y., and Heemink, A.: Machine learning for observation bias correction with application to dust storm data assimilation, Atmos. Chem. Phys., 19, 10009-10026, https://doi.org/10.5194/acp-19-10009-2019, 2019.

Kong, L., Tang, X., Zhu, J., Wang, Z., Liu, B., Zhu, Y., Zhu, L., Chen, D., Hu, K., Wu, H., Wu, Q., Shen, J., Sun, Y., Liu, Z., Xin, J., Ji, D., and Zheng, M.: High-resolution Simulation Dataset of Hourly $PM_{2.5}$ Chemical Composition in China (CAQRA-aerosol) from 2013 to 2020, Adv. Atmos. Sci., 42, 697-712, https://doi.org/10.1007/s00376-024-4046-5, 2025.

Kong, L., Tang, X., Zhu, J., Wang, Z., Li, J., Wu, H., Wu, Q., Chen, H., Zhu, L., Wang, W., Liu, B., Wang, Q., Chen, D., Pan, Y., Song, T., Li, F., Zheng, H., Jia, G., Lu, M., Wu, L., and Carmichael, G. R.: A 6-year-long (2013-2018) high-resolution air quality reanalysis dataset in China based on the assimilation of surface observations from CNEMC, Earth Syst. Sci. Data, 13, 529-570, https://doi.org/10.5194/essd-13-529-2021, 2021.

Lai, Y.: Application and Effectiveness Evaluation of Bayesian Optimization Algorithm in Hyperparameter Tuning of Machine Learning Models, in: 2024 International Conference on Power, Electrical Engineering, Electronics and Control (PEEEC), Athens, Greece, 14-16 August 2024, 351-355, https://doi.org/10.1109/PEEEC63877.2024.00070, 2024.

Lee, S., Park, S., Lee, M.-I., Kim, G., Im, J., and Song, C.-K.: Air Quality Forecasts Improved by Combining Data Assimilation and Machine Learning With Satellite AOD, Geophys. Res. Lett., 49, e2021GL096066, https://doi.org/10.1029/2021GL096066, 2022.

Legler, S. and Janjić, T.: Combining data assimilation and machine learning to estimate parameters of a convective-scale model, Q. J. R. Meteorol. Soc., 148, 860-874, https://doi.org/10.1002/qj.4235, 2022.

Lei, L. and Whitaker, J. S.: Evaluating the trade-offs between ensemble size and ensemble resolution in an ensemble-variational data assimilation system, J. Adv. Model Earth Sy., 9, 781-789, https://doi.org/10.1002/2016MS000864, 2017.

Lei, L., Sun, Y., Ouyang, B., Qiu, Y., Xie, C., Tang, G., Zhou, W., He, Y., Wang, Q., Cheng, X., Fu, P., and Wang, Z.: Vertical Distributions of Primary and Secondary Aerosols in Urban Boundary Layer: Insights into Sources, Chemistry, and Interaction with Meteorology, Environ. Sci. Technol., 55, 4542-4552, https://doi.org/10.1021/acs.est.1c00479, 2021.

Li, H. and Yang, T.: OIRF-LEnKF v1.0, Zenodo [code and data set], https://doi.org/10.5281/zenodo.16735735, 2025.

Li, H., Yang, T., Du, Y., Tan, Y., and Wang, Z.: Interpreting hourly mass concentrations of $PM_{2.5}$ chemical components with an optimal deep-learning model, J. Environ. Sci., 151, 125-139, https://doi.org/10.1016/j.jes.2024.03.037, 2025.

Li, H., Yang, T., Nerger, L., Zhang, D., Zhang, D., Tang, G., Wang, H., Sun, Y., Fu, P., Su, H., and Wang, Z.: NAQPMS-PDAF v2.0: a novel hybrid nonlinear data assimilation system for improved simulation of $PM_{2.5}$ chemical components,

Geosci. Model Dev., 17, 8495-8519, https://doi.org/10.5194/gmd-17-8495-2024, 2024.

Li, J., Wang, Y., Steenland, K., Liu, P., van Donkelaar, A., Martin, R. V., Chang, H. H., Caudle, W. M., Schwartz, J., Koutrakis, P., and Shi, L.: Long-term effects of $PM_{2.5}$ components on incident dementia in the northeastern United States, Innovation (Cambridge (Mass.)), 3, 100208-100208, https://doi.org/10.1016/j.xinn.2022.100208, 2022.

Lin, H., Jin, J., and van den Herik, J.: Air Quality Forecast through Integrated Data Assimilation and Machine Learning, in: Proceedings of the 11th International Conference on Agents and Artificial Intelligence, Prague, Czech Republic, 2, 787-793, https://doi.org/10.5220/0007555207870793, 2019.

Liu, S., Geng, G., Xiao, Q., Zheng, Y., Liu, X., Cheng, J., and Zhang, Q.: Tracking Daily Concentrations of $PM_{2.5}$ Chemical Composition in China since 2000, Environ. Sci. Technol., 56, 16517-16527, https://doi.org/10.1021/acs.est.2c06510, 2022.

Luo, Z., Han, Y., Hua, K., Zhang, Y., Wu, J., Bi, X., Dai, Q., Liu, B., Chen, Y., Long, X., and Feng, Y.: The effect of emission source chemical profiles on simulated $PM_{2.5}$ components: sensitivity analysis with the Community Multiscale Air Quality (CMAQ) modeling system version 5.0.2, Geosci. Model Dev., 16, 6757-6771, https://doi.org/10.5194/gmd-16-6757-2023, 2023.

Mallet, V. and Sportisse, B.: Uncertainty in a chemistry-transport model due to physical parameterizations and numerical approximations: An ensemble approach applied to ozone modeling, J. Geophys. Res.: Atmospheres, 111, https://doi.org/10.1029/2005jd006149, 2006.

Miao, R., Chen, Q., Zheng, Y., Cheng, X., Sun, Y., Palmer, P. I., Shrivastava, M., Guo, J., Zhang, Q., Liu, Y., Tan, Z., Ma, X., Chen, S., Zeng, L., Lu, K., and Zhang, Y.: Model bias in simulating major chemical components of $PM_{2.5}$ in China, Atmos. Chem. Phys., 20, 12265-12284, https://doi.org/10.5194/acp-20-12265-2020, 2020.

Nenes, A., Pandis, S. N., and Pilinis, C.: ISORROPIA: A new thermodynamic equilibrium model for multiphase multicomponent inorganic aerosols, Aquat. Geochem., 4, 123-152, https://doi.org/10.1023/A:1009604003981, 1998.

Nerger, L., Janjić, T., Schröter, J., and Hiller, W.: A regulated localization scheme for ensemble-based Kalman filters, Q.J.R. Meteorol. Soc., 138, 802-812, https://doi.org/10.1002/qj.945, 2012.

Probst, P., Wright, M. N., and Boulesteix, A.-L.: Hyperparameters and tuning strategies for random forest, WIREs Data Mining Knowl. Discov., 9, e1301, https://doi.org/10.1002/widm.1301, 2019.

Randles, C. A., da Silva, A. M., Buchard, V., Colarco, P. R., Darmenov, A., Govindaraju, R., Smirnov, A., Holben, B., Ferrare, R., Hair, J., Shinozuka, Y., and Flynn, C. J.: The MERRA-2 aerosol reanalysis, 1980 onward. Part I: System description and data assimilation evaluation, J. Clim., 30, 6823-6850, https://doi.org/10.1175/JCLI-D-16-0609.1, 2017.

Rasmussen, C. E.: Gaussian processes in machine learning, in: Advanced Lectures on Machine Learning, edited by: Bousquet, O., von Luxburg, U., Rätsch, G., Springer, Berlin, Heidelberg, 63-71, https://doi.org/10.1007/978-3-540-28650-9_4, 2004.

Shaheen, K., Hanif, M. A., Hasan, O., and Shafique, M.: Continual Learning for Real-World Autonomous Systems: Algorithms, Challenges and Frameworks, J. Intell. Robot. Syst., 105, 9, https://doi.org/10.1007/s10846-022-01603-6, 2022.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and Freitas, N. d.: Taking the Human Out of the Loop: A Review of

Bayesian Optimization, Proceedings of the IEEE, 104, 148-175, https://doi.org/10.1109/JPROC.2015.2494218, 2016.

Shi, Y., Liu, L., Hu, F., Fan, G., and Huo, J.: Nocturnal Boundary Layer Evolution and Its Impacts on the Vertical Distributions of Pollutant Particulate Matter, Atmosphere, 12, 610, https://doi.org/10.3390/atmos12050610, 2021.

Soni, A., Mandariya, A. K., Rajeev, P., Izhar, S., Singh, G. K., Choudhary, V., Qadri, A. M., Gupta, A. D., Singh, A. K., and Gupta, T.: Multiple site ground-based evaluation of carbonaceous aerosol mass concentrations retrieved from CAMS and MERRA-2 over the Indo-Gangetic Plain, Environ. Sci.: Atmos., 1, 577-590, https://doi.org/10.1039/d1ea00067e, 2021.

Stier, P., van den Heever, S. C., Christensen, M. W., Gryspeerdt, E., Dagan, G., Saleeby, S. M., Bollasina, M., Donner, L., Emanuel, K., Ekman, A. M. L., Feingold, G., Field, P., Forster, P., Haywood, J., Kahn, R., Koren, I., Kummerow, C., L'Ecuyer, T., Lohmann, U., Ming, Y., Myhre, G., Quaas, J., Rosenfeld, D., Samset, B., Seifert, A., Stephens, G., and Tao, W.-K.: Multifaceted aerosol effects on precipitation, Nat. Geosci., 17, 719-732, https://doi.org/10.1038/s41561-024-01482-6, 2024.

Stockwell, W. R., Middleton, P., Chang, J. S., and Tang, X.: The second generation regional acid deposition model chemical mechanism for regional air quality modeling, J. Geophys. Res.: Atmospheres, 95, 16343-16367, https://doi.org/10.1029/JD095iD10p16343, 1990.

Sun, Y.: Vertical structures of physical and chemical properties of urban boundary layer and formation mechanisms of atmospheric pollution, Chinese Sci. Bull., 63, 1374-1389, https://doi.org/10.1360/n972018-00258, 2018.

Sun, Y. L., Wang, Z. F., Wild, O., Xu, W. Q., Chen, C., Fu, P. Q., Du, W., Zhou, L. B., Zhang, Q., and Han, T. T.: "APEC Blue": Secondary Aerosol Reductions from Emission Controls in Beijing, Sci. Rep., 6, 20668, https://doi.org/10.1038/srep20668, 2016.

Valler, V., Franke, J., and Brönnimann, S.: Impact of different estimations of the background-error covariance matrix on climate reconstructions based on data assimilation, Clim. Past, 15, 1427-1441, https://doi.org/10.5194/cp-15-1427-2019, 2019.

Wang, H. L., Qiao, L. P., Lou, S. R., Zhou, M., Ding, A. J., Huang, H. Y., Chen, J. M., Wang, Q., Tao, S. K., Chen, C. H., Li, L., and Huang, C.: Chemical composition of $PM_{2.5}$ and meteorological impact among three years in urban Shanghai, China, J. Clean. Prod., 112, 1302-1311, https://doi.org/10.1016/j.jclepro.2015.04.099, 2016.

Weagle, C. L., Snider, G., Li, C., van Donkelaar, A., Philip, S., Bissonnette, P., Burke, J., Jackson, J., Latimer, R., Stone, E., Abboud, I., Akoshile, C., Anh, N. X., Brook, J. R., Cohen, A., Dong, J., Gibson, M. D., Griffith, D., He, K. B., Holben, B. N., Kahn, R., Keller, C. A., Kim, J. S., Lagrosas, N., Lestari, P., Khian, Y. L., Liu, Y., Marais, E. A., Martins, J. V., Misra, A., Muliane, U., Pratiwi, R., Quel, E. J., Salam, A., Segev, L., Tripathi, S. N., Wang, C., Zhang, Q., Brauer, M., Rudich, Y., and Martin, R. V.: Global Sources of Fine Particulate Matter: Interpretation of $PM_{2.5}$ Chemical Composition Observed by SPARTAN using a Global Chemical Transport Model, Environ. Sci. Technol., 52, 11670-11681, https://doi.org/10.1021/acs.est.8b01658, 2018.

Wei, J., Li, Z., Chen, X., Li, C., Sun, Y., Wang, J., Lyapustin, A., Brasseur, G. P., Jiang, M., Sun, L., Wang, T., Jung, C. H., Qiu, B., Fang, C., Liu, X., Hao, J., Wang, Y., Zhan, M., Song, X., and Liu, Y.: Separating Daily 1 km $PM_{2.5}$ Inorganic

Chemical Composition in China since 2000 via Deep Learning Integrating Ground, Satellite, and Model Data, Environ. Sci. Technol., 57, 18282-18295, https://doi.org/10.1021/acs.est.3c00272, 2023.

Wu, C., Cao, C., Li, J., Lv, S., Li, J., Liu, X., Zhang, S., Liu, S., Zhang, F., Meng, J., and Wang, G.: Different physicochemical behaviors of nitrate and ammonium during transport: a case study on Mt. Hua, China, Atmos. Chem. Phys., 22, 15621-15635, https://doi.org/10.5194/acp-22-15621-2022, 2022.

Wu, J., Chen, X.-Y., Zhang, H., Xiong, L.-D., Lei, H., and Deng, S.-H.: Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimizationb, J. Electron. Sci. Technol., 17, 26-40, https://doi.org/10.11989/JEST.1674-862X.80904120, 2019.

Xie, X., Hu, J., Qin, M., Guo, S., Hu, M., Wang, H., Lou, S., Li, J., Sun, J., Li, X., Sheng, L., Zhu, J., Chen, G., Yin, J., Fu, W., Huang, C., and Zhang, Y.: Modeling particulate nitrate in China: Current findings and future directions, Environ. Int., 166, 107369, https://doi.org/10.1016/j.envint.2022.107369, 2022.

Yang, L. M. and Grooms, I.: Machine learning techniques to construct patched analog ensembles for data assimilation, J. Comput. Phys., 443, 110532, https://doi.org/10.1016/j.jcp.2021.110532, 2021.

Yang, T., Li, H., Xu, W., Song, Y., Xu, L., Wang, H., Wang, F., Sun, Y., Wang, Z., and Fu, P.: Strong Impacts of Regional Atmospheric Transport on the Vertical Distribution of Aerosol Ammonium over Beijing, Environ. Sci. Technol. Letters, 11, 29-34, https://doi.org/10.1021/acs.estlett.3c00791, 2024.

Zaveri, R. A. and Peters, L. K.: A new lumped structure photochemical mechanism for large-scale applications, J. Geophys. Res.: Atmospheres, 104, 30387-30415, https://doi.org/10.1029/1999JD900876, 1999.

Zhao, C., Sun, Y., Yang, J., Li, J., Zhou, Y., Yang, Y., Fan, H., and Zhao, X.: Observational evidence and mechanisms of aerosol effects on precipitation, Sci. Bull., 69, 1569-1580, https://doi.org/10.1016/j.scib.2024.03.014, 2024.