Geosci. Model Dev., 19, 4835–4856, 2026
https://doi.org/10.5194/gmd-19-4835-2026

Geoscientific Model Development
Open Access
EGU

OIRF-LEnKF v1.0: a novel data assimilation system by integrating incremental machine learning with a localized EnKF for enhanced $PM_{2.5}$ chemical component simulation and reanalysis

Hongyi Li[1], Ting Yang[1], Lei Kong[1], Di Zhang[2], Guigang Tang[2], Xiao Tang[1,3], and 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
[3]College of Earth and Planetary Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

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

Received: 13 August 2025 – Discussion started: 11 September 2025
Revised: 24 December 2025 – Accepted: 21 February 2026 – Published: 10 June 2026

Abstract. Assimilating observational data into numerical simulation 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 simulation skill, as it solely provides optimal initial conditions. This paper introduces an incrementally updatable machine learning-based data assimilation system (Optimized Incremental Random Forest coupled with Localized Ensemble Kalman Filter, OIRF-LEnKF v1.0) that achieves high efficiency and high quality in generating background and analysis fields for 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, primarily because the simulation process requires only 0.13 %–0.20 % of the CTM computation time. Sensitivity tests demonstrate that the incremental learning during the simulation process enhances the percentage change of the Pearson correlation coefficient relative to its minimum value (ΔCORR) by 2.43 %–11.75 % and reduces the percentage change of the RMSE relative to its maximum value (ΔRMSE) by 32.55 %–40.36 %, 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 and 2.36 $\mu g\,m^{-3}$ during the simulation 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 simulation process exhibit a significant reduction of 33.16 %–90.10 % compared to those of the CTM-based DA, highlighting the superior simulation 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 $\mu g\,m^{-3}$), the dataset generated by OIRF-LEnKF v1.0 (CORR: 0.97, RMSE: 1.12 $\mu g\,m^{-3}$) 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 background field (Geer, 2021). DA 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 limited (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., 2024a). However, CTM-based NP2 requires a reduction in ensemble size to maintain computational efficiency during simulation 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 background fields in NP2 to assimilation frequency suggests that improvements in initial conditions have limited effects on enhancing the simulation 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 generating background 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 background fields for DA (Dong et al., 2022, 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.

The Random Forest (RF) model (Gohari et al., 2025; Lin et al., 2022; Lv et al., 2021; Meng et al., 2018) and Deep Neural Networks (DNNs) (Li et al., 2025; Liu et al., 2023) have been widely used for simulating and predicting $PM_{2.5}$ chemical component concentrations, with DNNs achieving a marginally superior predictive accuracy. However, a single DNN is outperformed by a RF model in terms of the computational efficiency during both training and inference (Debjyoti and Utpal, 2025; Jalali et al., 2025; Xi, 2022). Within an ensemble DA framework, periodically creating and running an ensemble of DNNs imposes a significant computa-

tional burden in contrast to the RF model, which inherently provides an ensemble. Consequently, the RF model offers an optimal trade-off between predictive performance and computational demand, making it a practical and efficient choice for coupling with ensemble DA.

This study proposes an optimized incremental Random Forest (OIRF) model as a solution to the challenges of computational inefficiency and inadequate advancements in generating background and analysis fields within traditional CTM-based DA. The OIRF model is capable of providing a large number of background ensemble members at a reduced computational cost, which helps mitigate the underestimation of background 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 simulation. Then, the OIRF model is online coupled with the localized ensemble Kalman filter (LEnKF) algorithm to develop a novel data assimilation system (OIRF-LEnKF v1.0), which achieves a rapid iteration for high-quality simulation, 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 simulation 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 model with the localized ensemble Kalman filter (LEnKF) algorithm (Fig. 1). The ML-based OIRF ensemble model offers an effective alternative to conventional CTMs by promptly supplying background 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 background fields, minimizing interference from spurious correlations by implementing localization schemes, thereby generating high-accuracy analysis fields for incremental learning of the OIRF model. The online coupling of the OIRF model with the LEnKF algorithm facilitates the iterative execution of ensemble simulation, assimilation, and incremental learning at each time step. Consequently, the OIRF-LEnKF v1.0 is capable of generating high-quality background and analysis fields while simultaneously undergoing incremental learning.

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*: incremental learning of the OIRF model at time steps > 1. High-quality analysis fields at the last time step, along with the corresponding meteorological and anthropogenic input data, are employed to train a new ensemble of decision trees. The old decision trees, which exhibit poor simulation performance, are subsequently replaced with new decision trees to enhance the simulation accuracy and generalization ability of the OIRF model.
- *Step 3*: generating a background 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 background fields using the LEnKF algorithm.
- *Step 5*: scoring the simulation performance of ensemble decision trees in the OIRF model using mean absolute error (MAE) and screening out the decision trees with poor simulation performance based on a predefined threshold. 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 ($\boldsymbol{\theta}_n$) to facilitate feature random selection and sample bootstrapping. This approach enhances the diversity among DTs while maintaining the predictive capability of each DT (Breiman, 2001). Unlike conventional ensemble simulations that rely on multiple CTMs, RF can swiftly generate an ensemble of background

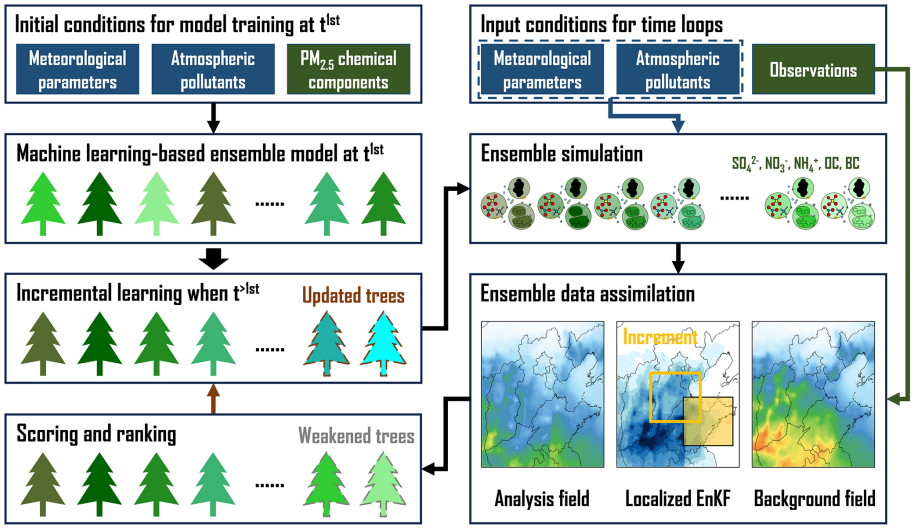

Figure 1. The framework of OIRF-LEnKF v1.0.

fields required for DA from multiple DTs without requiring external ensemble perturbation. The final simulation of the RF model is represented by the average of all DT outputs (Eq. 1).

$$f^{\mathrm{RF}}(\boldsymbol{x}) = \frac{1}{N}\sum_{n=1}^{N} f^{\mathrm{DT}}(\boldsymbol{x}, \boldsymbol{\theta}_n), \tag{1}$$

where $\boldsymbol{x}$ represents the input features, including anthropogenic atmospheric pollutants and meteorological conditions. $f^{\mathrm{RF}}(\boldsymbol{x})$ denotes the simulation of $PM_{2.5}$ chemical component concentrations. N is the total number of DTs. $f^{\mathrm{DT}}(\boldsymbol{x}, \boldsymbol{\theta}_n)$ denotes the output of the nth DT and $\boldsymbol{\theta}_n$ is an independently and identically distributed random vector that facilitates feature random selection and sample bootstrapping. The criterion for selecting the optimal split at each node during the training of an individual DT involves maximizing the reduction in mean squared error (MSE) over all splitting candidates.

Inspired by the idea of dynamically updating DTs with weak performance (Xie et al., 2016), the OIRF model incorporates a novel incremental learning mechanism into the RF model, enabling it to conduct effective updating from newly available training data within a simulation-assimilation cycle. In the incremental learning mechanism, the OIRF model scores the simulation performance of each DT based on the mean absolute error (MAE), as shown in Eq. (2). The MAE is quantified by the DT outputs and high-accuracy analysis fields at the same time step. A leakage-aware evaluation indicates that using the analysis field as scoring target did not cause substantial information leakage, while employing the independent high-quality observation as scoring target is also recommended (Sect. S1 in the Supplement).

$$f_n^{\mathrm{score}} = \frac{1}{K}\sum_{i=1}^{K}\left|x_i^{\mathrm{ana}} - f^{\mathrm{DT}}(x_i, \boldsymbol{\theta}_n)\right|, n = 1, 2, \ldots, N \tag{2}$$

Here, f_n^{score} is the MAE value of the nth DT. K is the total number of grid points of $PM_{2.5}$ chemical component concentrations. x_i^{ana} is the analysis value of concentrations at the ith grid point after DA. $f^{\mathrm{DT}}(x_i, \boldsymbol{\theta}_n)$ denotes the simulation value of the nth DT at the ith grid point. Notably, x_i used in machine learning denotes the input features, while x_i^{ana} used in data assimilation denotes the analysis states.

The incremental learning mechanism introduces a threshold (τ_p) to screen out the DTs with poor simulation performance. The threshold is defined as the pth percentile value of f_n^{score}. The percentile-based threshold ensures a stable and controllable number of DTs are updated, a critical feature for maintaining the smoothness and stability of the estimation of background error covariance within the ensemble data assimilation framework and preventing model overfitting to the new information. 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.

$$f_t^{\mathrm{DT}} = \begin{cases} f^{\mathrm{DT}}\left(\boldsymbol{x}, \boldsymbol{\theta}_n | \boldsymbol{x}_{t-\Delta t}^{\mathrm{ana}}\right), f_n^{\mathrm{score}} \le \tau_p, & n = 1, 2, \ldots, N_p \\ f^{\mathrm{DT}}\left(\boldsymbol{x}, \boldsymbol{\theta}_n | \boldsymbol{x}_t^{\mathrm{ana}}\right), f_n^{\mathrm{score}} > \tau_p, & n = N_p + 1, N_p + 2, \ldots, N \end{cases} \tag{3}$$

Here, f_t^{DT} represents the final output of the updated DTs following incremental learning at time t. $f^{\mathrm{DT}}(\boldsymbol{x}, \boldsymbol{\theta}_n | \boldsymbol{x}_{t-\Delta t}^{\mathrm{ana}})$ denotes the output of the retained old DTs while $f^{\mathrm{DT}}(\boldsymbol{x}, \boldsymbol{\theta}_n | \boldsymbol{x}_t^{\mathrm{ana}})$ refers to the output of the new DTs. Δt represents the time interval of incremental learning. τ_p indicates the pth percentile value of $f_n^{\mathrm{score}} (n = 1, 2, \ldots, N)$, and 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 and artificially optimistic performance into ensemble simulation of the OIRF model.

The final simulation ($f^{\mathrm{OIRF}}(\boldsymbol{x})$) of the OIRF model at time t is derived from Eq. (4) by averaging the outputs of the updated DTs.

$$f_t^{\mathrm{OIRF}}(\boldsymbol{x}) = \frac{1}{N}\sum_{n=1}^{N} f_t^{\mathrm{DT}}(\boldsymbol{x}, \boldsymbol{\theta}_n) \tag{4}$$

Notably, the incremental learning mechanism generates new DTs within a Bayesian optimization framework, which ensures that the updated RF model simultaneously acquires new knowledge and preserves optimal hyperparameters over time. Consequently, the incremental learning mechanism enhances the capacity of the OIRF model to incorporate newly available training data and replace the underperforming DTs with deterministically superior ones, thereby dynamically improving its generalization ability in simulating $PM_{2.5}$ chemical component concentrations.

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). The objective function was defined by Eq. (5). 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 nonparametric Gaussian process regression model (Rasmussen, 2004, 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).

$$J(\theta) = \ln\left(1 + \frac{1}{N}\sum_{i=1}^{N}\left(y_i^{\mathrm{pred}}(\theta) - y_i^{\mathrm{o}}\right)^2\right) \tag{5}$$

Here, $J(\theta)$ represents the objective value, θ represents the set of hyperparameters under optimization, N is the total number of samples in the training dataset. $y_i^{\mathrm{pred}}(\theta)$ is the predicted value for the ith sample, y_i^{o} is the observation value for the ith sample.

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, 2003). In this study, the OIRF model replaced the conventional CTMs to provide an ensemble of DT-simulated background fields for estimating the background error covariance (Eq. 6). The ensemble size in DA is equal to the total number of DTs in the OIRF model.

$$\begin{aligned}\mathbf{P}_t^{\mathrm{f}} =& \frac{1}{N-1}\sum_{n=1}^{N}\left(f_t^{\mathrm{DT}}(\boldsymbol{x}, \boldsymbol{\theta}_n) - \overline{f_t^{\mathrm{DT}}(\boldsymbol{x}, \boldsymbol{\theta}_n)}\right) \\ & \left(f_t^{\mathrm{DT}}(\boldsymbol{x}, \boldsymbol{\theta}_n) - \overline{f_t^{\mathrm{DT}}(\boldsymbol{x}, \boldsymbol{\theta}_n)}\right)^T\end{aligned} \tag{6}$$

Here, $\mathbf{P}_t^{\mathrm{f}}$ is the flow-dependent background error covariance matrix of $PM_{2.5}$ chemical component concentrations at time t, $\overline{f_t^{\mathrm{DT}}(\boldsymbol{x}, \boldsymbol{\theta}_n)}$ refers to the ensemble mean across decision trees in the random forest at time t.

The Kalman gain matrix ($\mathbf{K}$) can be calculated by Eqs. (7)–(9).

$$\mathbf{K} = \mathbf{P}_t^{\mathrm{f}}\mathbf{H}_t^T\left(\mathbf{H}_t\mathbf{P}_t^{\mathrm{f}}\mathbf{H}_t^T + \mathbf{R}_t\right)^{-1}, \tag{7}$$

$$\begin{aligned}\mathbf{P}_t^{\mathrm{f}}\mathbf{H}_t^T =& \frac{1}{N-1}\sum_{n=1}^{N}\left(f_t^{\mathrm{DT}}(\boldsymbol{x}, \boldsymbol{\theta}_n) - \overline{f_t^{\mathrm{DT}}(\boldsymbol{x}, \boldsymbol{\theta}_n)}\right) \\ & \left(H\left(f_t^{\mathrm{DT}}(\boldsymbol{x}, \boldsymbol{\theta}_n)\right) - \overline{H\left(f_t^{\mathrm{DT}}(\boldsymbol{x}, \boldsymbol{\theta}_n)\right)}\right)^T,\end{aligned} \tag{8}$$

$$\begin{aligned}\mathbf{H}_t\mathbf{P}_t^{\mathrm{f}}\mathbf{H}_t^T =& \frac{1}{N-1}\sum_{n=1}^{N}\left(H\left(f_t^{\mathrm{DT}}(\boldsymbol{x}, \boldsymbol{\theta}_n)\right) - \overline{H\left(f_t^{\mathrm{DT}}(\boldsymbol{x}, \boldsymbol{\theta}_n)\right)}\right) \\ & \left(H\left(f_t^{\mathrm{DT}}(\boldsymbol{x}, \boldsymbol{\theta}_n)\right) - \overline{H\left(f_t^{\mathrm{DT}}(\boldsymbol{x}, \boldsymbol{\theta}_n)\right)}\right)^T,\end{aligned} \tag{9}$$

Here, $\mathbf{K}$ is the Kalman gain matrix. $\mathbf{H}_t$ is the observation operator at t. $\mathbf{R}_t$ 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

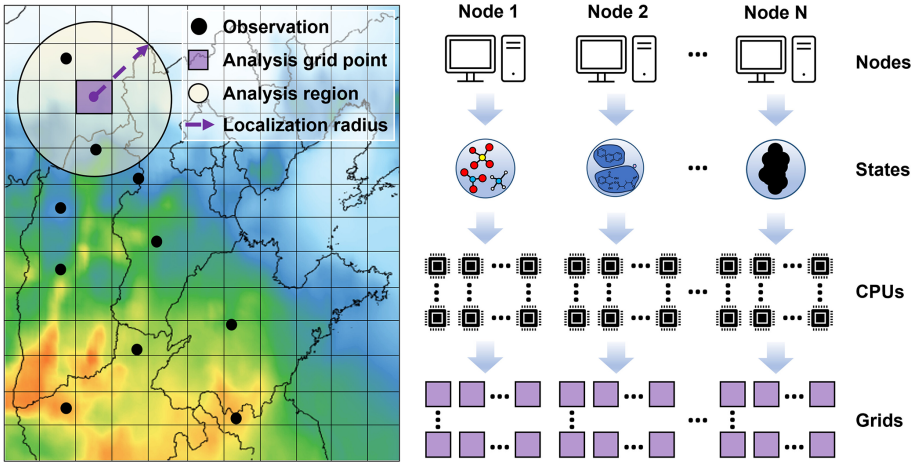

Figure 2. The scheme for domain localization and parallelization.

conducts spatial mapping between the observations and the background fields due to consistency in the variable and temporal dimensions. The method employed for spatial mapping between observations from sparse sites and gridded background fields is the k-nearest neighbor search (Friedman et al., 1977).

The final analysis fields ($\boldsymbol{x}_{n,t}^{\mathrm{ana}}$) can be obtained from the integration of background fields ($f_t^{\mathrm{DT}}(x, \boldsymbol{\theta}_n)$) and observations ($y_t^{\mathrm{o}}$):

$$\boldsymbol{x}_{n,t}^{\mathrm{ana}} = f_t^{\mathrm{DT}}(\boldsymbol{x}, \boldsymbol{\theta}_n) + \mathbf{K}\left(\boldsymbol{y}_t^{\mathrm{o}} + y'^{\mathrm{o}}_{n,t} - H\left(f_t^{\mathrm{DT}}(\boldsymbol{x}, \boldsymbol{\theta}_n)\right)\right), \quad n = 1, 2, \ldots, N \tag{10}$$

Here, $\boldsymbol{x}_{n,t}^{\mathrm{ana}}$ is the analysis field of the nth ensemble member at t. $\boldsymbol{y}_t^{\mathrm{o}}$ is the observation of $PM_{2.5}$ chemical components at t and $y'^{\mathrm{o}}_{n,t}$ is the observation perturbation of the nth 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 for each analysis grid point using only background fields and observations within a specific localization radius (Fig. 2), with the same update form as global EnKF (Eq. 10). The fundamental update form is presented in Eq. (11).

$$\boldsymbol{x}_{n,\delta}^{\mathrm{ana}} = f_\delta^{\mathrm{DT}}(\boldsymbol{x}, \boldsymbol{\theta}_n) + \mathbf{K}_\delta\left(\boldsymbol{y}_\delta^{\mathrm{o}} + y'^{\mathrm{o}}_{n,\delta} - H_\delta\left(f_\delta^{\mathrm{DT}}(\boldsymbol{x}, \boldsymbol{\theta}_n)\right)\right), \quad n = 1, 2, \ldots, N \tag{11}$$

Here, $\boldsymbol{x}_{n,\delta}^{\mathrm{ana}}$ is the analysis value within the localization domain δ of the nth ensemble member. $f_\delta^{\mathrm{DT}}(\boldsymbol{x}, \boldsymbol{\theta}_n)$ is the background value within the localization domain δ of the nth ensemble member. $\mathbf{K}_\delta$ is the local Kalman gain matrix computed from the ensemble covariance within the localization domain δ. $\boldsymbol{y}_\delta^{\mathrm{o}}$ is the observation of $PM_{2.5}$ chemical components within the localization domain δ and $y'^{\mathrm{o}}_{n,\delta}$ is the observation perturbation of the nth ensemble member within the localization domain δ. H_δ is the linear observation operator within the localization domain δ.

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 and eliminate the need for inter-node communication. Subsequently, the grid points of each chemical component are assigned to multiple CPUs within these independent computational nodes.

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 ($\mathbf{R}_t$) and a distance-based weight matrix ($\mathbf{W}$) as shown in Eq. (12).

$$\mathbf{K}^{\mathrm{L}} = \mathbf{P}_t^{\mathrm{f}}\mathbf{H}_t^T\left(\mathbf{H}_t\mathbf{P}_t^{\mathrm{f}}\mathbf{H}_t^T + \mathbf{W}\cdot\mathbf{R}_t\right)^{-1} \quad (12)$$

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

The distance-based weight matrix ($\mathbf{W}_i$) for the ith localization domain is obtained using a Gaussian function:

$$\mathbf{W}_i = \mathrm{diag}\left(\exp\left(\frac{-d(i,j)^2}{2L^2}\right)\right) j = 1, 2, \ldots, N_{\mathrm{obs}}. \quad (13)$$

Here, $d(i, j)$ is the Euclidean distance between center grid point of the ith localization domain and observation point j. L is the decorrelation length. N_{obs} is the total number of effective observations within the ith localization domain. $\mathbf{W}$ is constructed as a diagonal matrix ($N_{\mathrm{obs}} \times N_{\mathrm{obs}}$), applying a distance-dependent weighting directly to the diagonal elements of observation error covariance matrix $\mathbf{R}_t$.

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 h, respectively. The data of the input feature utilized for training the OIRF model are outlined in Sect. 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 h intervals, 12 h intervals, 6 h intervals, and 1 h intervals. The experimental design is detailed in Sect. 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 randomly 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 and 80 km, respectively, based on the spatial range and resolution requirements. The assimilation frequency matches the temporal resolution of 1 h.

2.1.5 Data

2.1.6 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, Tang et al., 2020). 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://doi.org/10.24381/cds.bd0915c6, Hersbach et al., 2023) with a horizontal resolution of 0.25° × 0.25° (Hersbach et al., 2023). 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/10.1007/s00376-024-4046-5, Kong et al., 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).

2.1.7 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., 2024a).

2.1.8 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://doi.org/10.24381/d58bbf47, Copernicus Atmosphere Monitoring Service, 2020), the Modern-Era Retrospective analysis for Research and Applications,

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

Category	Parameter	Setting
Ensemble simulation	State variable	SO_4^{2-}, NO_3^-, NH_4^+, OC and BC
	Model domain	North China (32.38–44.90° N, 108.07–127.01° 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	State dimension	5, including SO_4^{2-}, NO_3^-, NH_4^+, OC and BC
	Latitudinal dimension Longitudinal dimension	249 grid points 300 grid points
	Algorithm	LEnKF
	Localization radius	200 km
	Decorrelation length	80 km
	Assimilation frequency	1 h

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, Li et al., 2024b). The High-resolution and High-quality Air Pollutants dataset for China (CHAP, https://doi.org/10.5281/zenodo.10011898, Wei et al., 2022) 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.

2.2 Experimental setting

We designed four experiments to evaluate the performance of OIRF-LEnKF v1.0 on background 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, simulation, 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 simulation 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 simulation and assimilation performance to ensemble size and update frequency. In the third experiment, we conducted a 2-month simulation-assimilation loop 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.

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	February 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°

3 Results and discussion

3.1 Computational efficiency

As shown in Fig. 3, we evaluate the computational efficiencies of hyperparameter tuning, simulation 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 Fig. 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.

The computational costs of OIRF-LEnKF v1.0 in simulation 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 simulation and assimilation for OIRF-LEnKF v1.0 amounts to only 11.41 % to 16.60 % of that for NP2, especially during the simulation process, which accounts for merely 0.13 % to 0.20 % (Fig. 3d). The marked improvement in simulation efficiency by OIRF-LEnKF v1.0 is comparable to the deep neural network model (Adie et al., 2024). This enhancement is primarily attributed to the fact that ML-based simulation does not necessitate a profound understanding of the complex physicochemical mechanisms of the atmosphere (Fang et al., 2022), whereas CTM-based simulation 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 and 39.53 s, respectively. Specifically, the time consumed by simulation increases by 0.22 and 39.53 s, respectively, while the time consumed by assimilation increases by 17.69 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 simulation 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 simulation and reanalysis 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 simulation at the simulation step. The update frequency for incremental learning drives the adaptability of the ML model to non-stationary data distributions (Shaheen et al., 2022), thereby influencing the generalization ability at the simulation step and indirectly

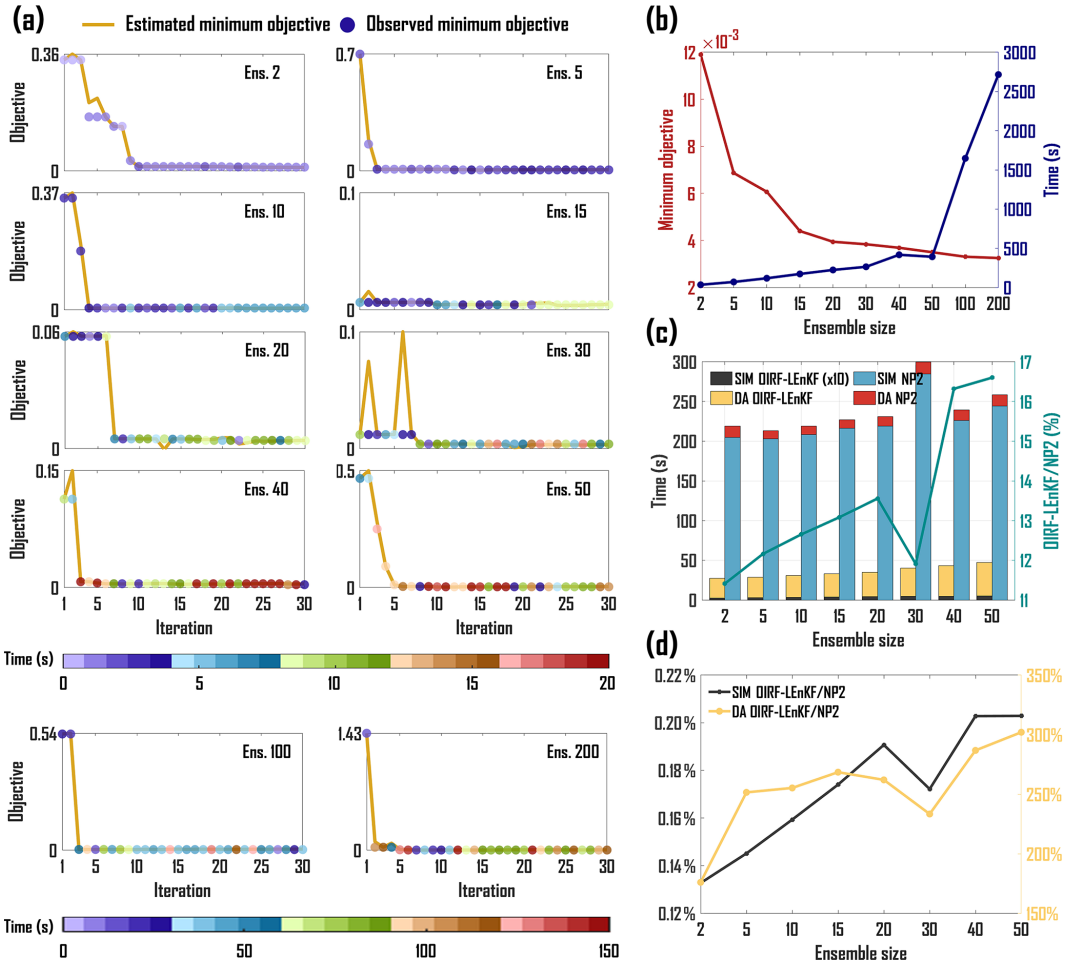

Figure 3. Computational efficiency of OIRF-LEnKF v1.0. **(a)** Variation in the minimum objective value throughout the Bayesian optimization process and time consumed by each iteration, determined by Eq. (5). **(b)** Minimum value of total observed minimum objectives and total time consumed during Bayesian optimization process for different ensemble sizes, **(c)** time consumed by model simulation 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 simulation and data assimilation between OIRF-LEnKF and NP2. SIM represents the simulation phase, and DA represents the data assimilation phase. The elapsed time of the OIRF-LEnKF simulation process in **(c)** has been magnified by a factor of 10 for better clarity.

affecting the background error information at the analysis step.

During the ML simulation process, the statistical indicators that compare the background 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 percentage change of CORR relative to minimum CORR (ΔCORR) rises by 2.43 % (ensembles size is 200) to 11.75 % (ensembles size is 30), and the percentage change of RMSE relative to maximum RMSE (ΔRMSE) decreases by 32.55 % (ensembles size is 20) to 40.36 % (ensembles size is 100) when comparing a 1 h 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 simulation accuracy in rapidly changing chemical component simulations. Notably, an increase in ensemble size can amplify the effect of incremental learning on simulation errors. Specifically, the reduction in ΔRMSE at an ensemble size of 100 is approximately 8 % greater than at an ensemble size of 20 when comparing a 1 h update frequency to a scenario without incremental learning (Fig. 4b), which is attributed to the fact that as the ensemble size increases, the probability density distribution be-

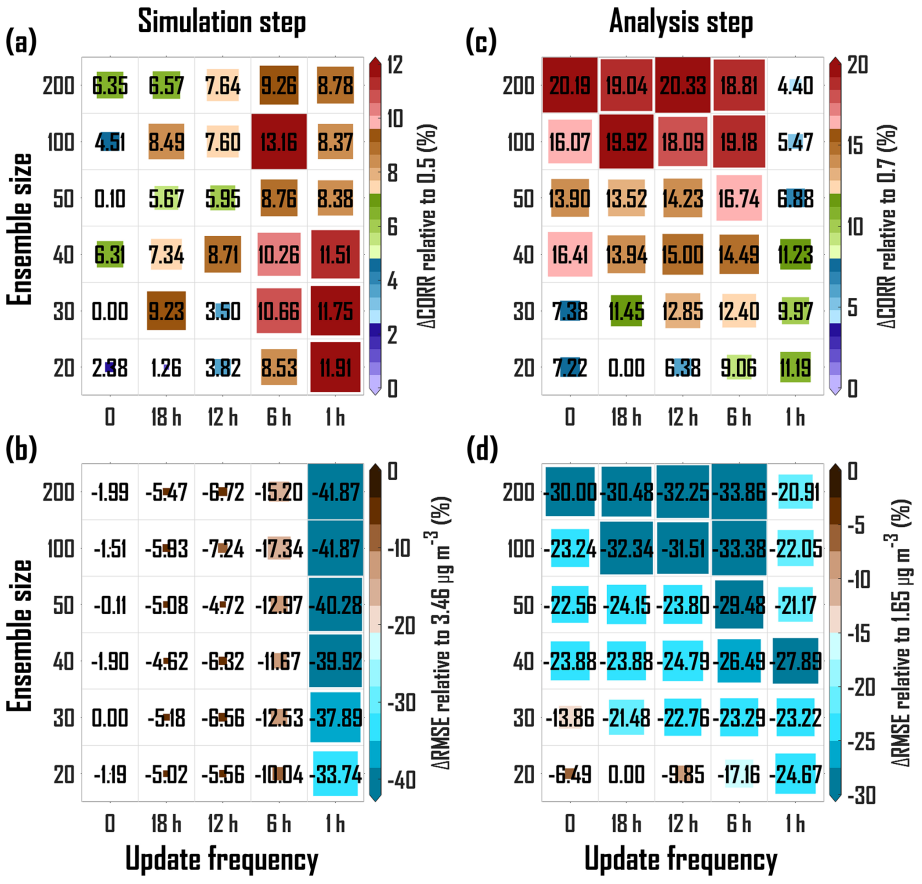

Figure 4. **(a)** Percentage change of Pearson correlation coefficient (CORR) relative to the minimum CORR (0.5) (ΔCORR, %) for sensitivity test with six ensemble sizes (20, 30, 40, 50, 100, 200) and five update frequencies (no update, 18 h interval, 12 h interval, 6 h interval and 1 h interval) at the simulation step. **(b)** Same as **(a)** but for percentage change of root mean square error (RMSE) relative to the maximum RMSE (3.46 μg m^{-3}) (ΔRMSE, %) at the simulation step. **(c)** Same as **(a)** but for percentage change of CORR relative to the minimum CORR (0.7) at the analysis step. **(d)** Same as **(a)** but for percentage change of RMSE relative to the maximum RMSE (1.65 μg m^{-3}) at the analysis step.

comes more accurate, leading to improved ensemble simulation skill (Chen, 2024).

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 h 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 ΔCORR increased by 9.75 % (update frequency is 6 h) to 19.04 % (update frequency is 18 h), and the ΔRMSE decreased by 16.70 % (update frequency is 6 h) to 30.48 % (update frequency is 18 h) 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 h 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 h update frequency can significantly enhance the simulation performance, we configured an ensemble size of 50 with a 6 h update frequency in OIRF-LEnKF v1.0 to balance computational efficiency, ML simulation accuracy, and DA analysis performance.

3.3 Evaluation of DA results

This section assesses the performance of the free-run field without DA and incremental learning (FR), the ML-simulated background field with incremental learning (SIM) 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 FR, SIM, and ANA across 33 ground-level sites. As illustrated in Fig. 5a1–a3, the errors of FR for NH_4^+, NO_3^-, and SO_4^{2-} ranged from $-2.30 \pm 1.97\,\mu g\,m^{-3}$ to $8.84 \pm 5.04\,\mu g\,m^{-3}$, $-7.60 \pm 5.29\,\mu g\,m^{-3}$ to $14.64 \pm 17.20\,\mu g\,m^{-3}$, and $-4.31 \pm 3.81\,\mu g\,m^{-3}$ to $9.61 \pm 6.00\,\mu g\,m^{-3}$, respectively. The overall errors of FR 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 SIM concentrated to a range of $-2.66 \pm 4.18\,\mu g\,m^{-3}$ to $5.18 \pm 4.87\,\mu g\,m^{-3}$ (NH_4^+), $-7.17 \pm 10.75\,\mu g\,m^{-3}$ to $10.07 \pm 7.48\,\mu g\,m^{-3}$ (NO_3^-), and $-1.37 \pm 1.98\,\mu g\,m^{-3}$ to $6.50 \pm 4.81\,\mu g\,m^{-3}$ (SO_4^{2-}), indicating that incremental learning enhances the ability to capture the temporal features of inorganic salt concentrations. Compared to FR and SIM, 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 FR for OC and BC ranged from $-12.18 \pm 4.09\,\mu g\,m^{-3}$ to $-1.11 \pm 2.78\,\mu g\,m^{-3}$ and $-5.41 \pm 1.39\,\mu g\,m^{-3}$ to $-0.87 \pm 0.57\,\mu g\,m^{-3}$, respectively, with a general overestimation of carbonaceous aerosol concentrations (Fig. 5a4 and a5). The errors of SIM and ANA are relatively similar, both concentrating around zero over time due to the effects of incremental learning and DA.

Figure 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 FR 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 $9.15\,\mu g\,m^{-3}$, 4.73 to $16.24\,\mu g\,m^{-3}$, 2.31 to $10.24\,\mu g\,m^{-3}$, 4.57 to $10.41\,\mu g\,m^{-3}$, and 1.36 to $3.42\,\mu g\,m^{-3}$, respectively (Fig. 5c). Following incremental learning, the CORR and RMSE values of SIM demonstrated a more concentrated data distribution than those of FR, with average CORR (0.42 to 0.83) and RMSE (0.99 to $7.80\,\mu g\,m^{-3}$) values increasing by 5.61 % to 114.28 % and decreasing by 26.38 % to 61.75 %, respectively. Additionally, compared to the SIM of a CTM-based DA system, the SIM 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 incremental learning mechanism is more effective than the optimal estimation of initial conditions in enhancing $PM_{2.5}$ chemical component simulations, which is attributed to the fact that the enhancement in ML-based simulation by incremental learning is global, while the CTM-based simulation 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 FR and SIM. The average CORR (0.58 to 1.00) and RMSE (0.80 to $2.36\,\mu g\,m^{-3}$) values demonstrated advancements of 35.27 % to 187.15 % and 68.99 % to 91.31 %, respectively, compared to FR, and advancements of 18.85 % to 38.73 % and 19.71 % to 88.20 %, respectively, compared to SIM.

For the VE sites without DA, the CORR values of FR 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 $8.25\,\mu g\,m^{-3}$, 8.04 to $14.18\,\mu g\,m^{-3}$, 3.94 to $7.04\,\mu g\,m^{-3}$, 6.23 to $10.05\,\mu g\,m^{-3}$, and 2.33 to $3.30\,\mu g\,m^{-3}$, respectively (Fig. 5e). After incremental learning, the CORR and RMSE values of SIM exhibited a more concentrated data distribution than those of FR, with average CORR (0.39 to 0.81) and RMSE (0.93 to $7.76\,\mu g\,m^{-3}$) values increasing by 12.00 % to 124.69 % and decreasing by 28.37 % to 68.00 %, respectively. Furthermore, compared to the SIM of a CTM-based DA system, the SIM 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 incremental learning mechanism for improving ML-based simulations 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 $7.76\,\mu g\,m^{-3}$, respectively, showing a more concentrated data distribution than those of FR and SIM. The average CORR and RMSE values increased by 14.14 % to 116.65 % and decreased by 23.46 % to 68.75 %, respectively, compared to FR, 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), FR, SIM 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 $12.20\,\mu g\,m^{-3}$ (Fig. 6a1). However, FR 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 and $3.07\,\mu g\,m^{-3}$, respectively (Fig. 7a1). This finding is attributed to the underestimation of the original training samples (Kong et al., 2025). Compared to FR, the SIM mitigates the underestimation (Fig. 6a3), with

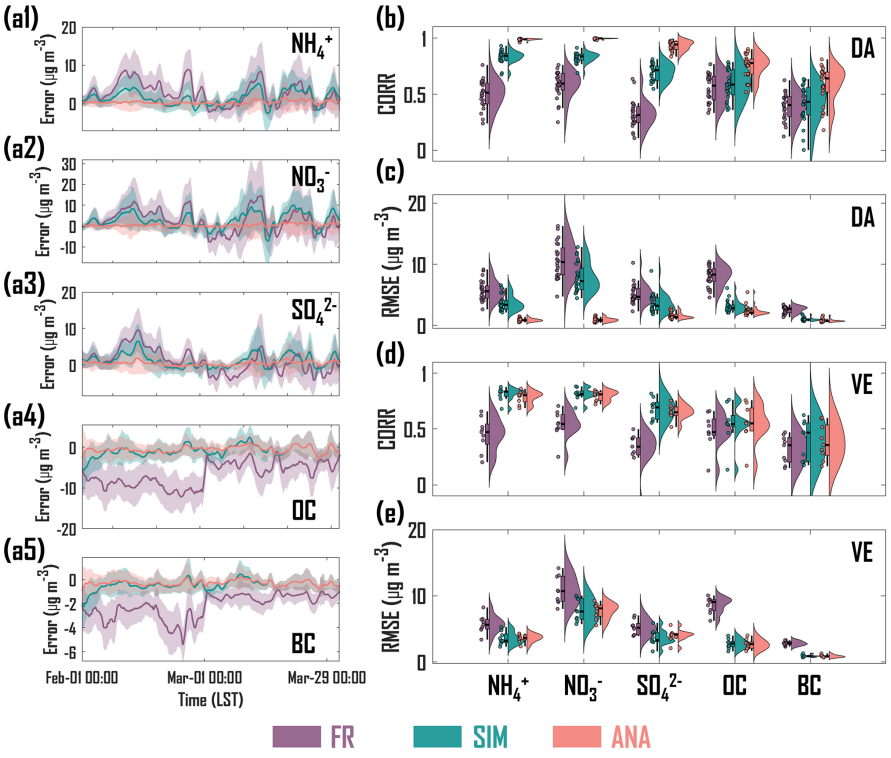

Figure 5. Smoothed variation in the error between observation and model output – including the free-run field (FR), the ML-simulated background field (SIM) and the analysis field (ANA) – for **(a1)** NH_4^+, **(a2)** NO_3^-, **(a3)** SO_4^{2-}, **(a4)** OC and **(a5)** BC at total sites during February 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.

Table 3. The correlation coefficient (CORR) and root mean square error (RMSE, $\mu g\,m^{-3}$) of OIRF-LEnKF v1.0 (this study) and NAQPMS-PDAF v2.0 (NP2) at DA sites and VE sites for the simulations of NH_4^+, NO_3^-, SO_4^{2-}, 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 ($\mu g\,m^{-3}$)										
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

96 % of DA sites underestimating by 1.56 $\mu g\,m^{-3}$ and 78 % of VE sites underestimating by 1.88 $\mu g\,m^{-3}$ (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 $\mu g\,m^{-3}$ and 44 % of VE sites underestimating by 2.34 $\mu g\,m^{-3}$, respectively (Fig. 7a3). The increment field (INC) between ANA and SIM 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 ad-

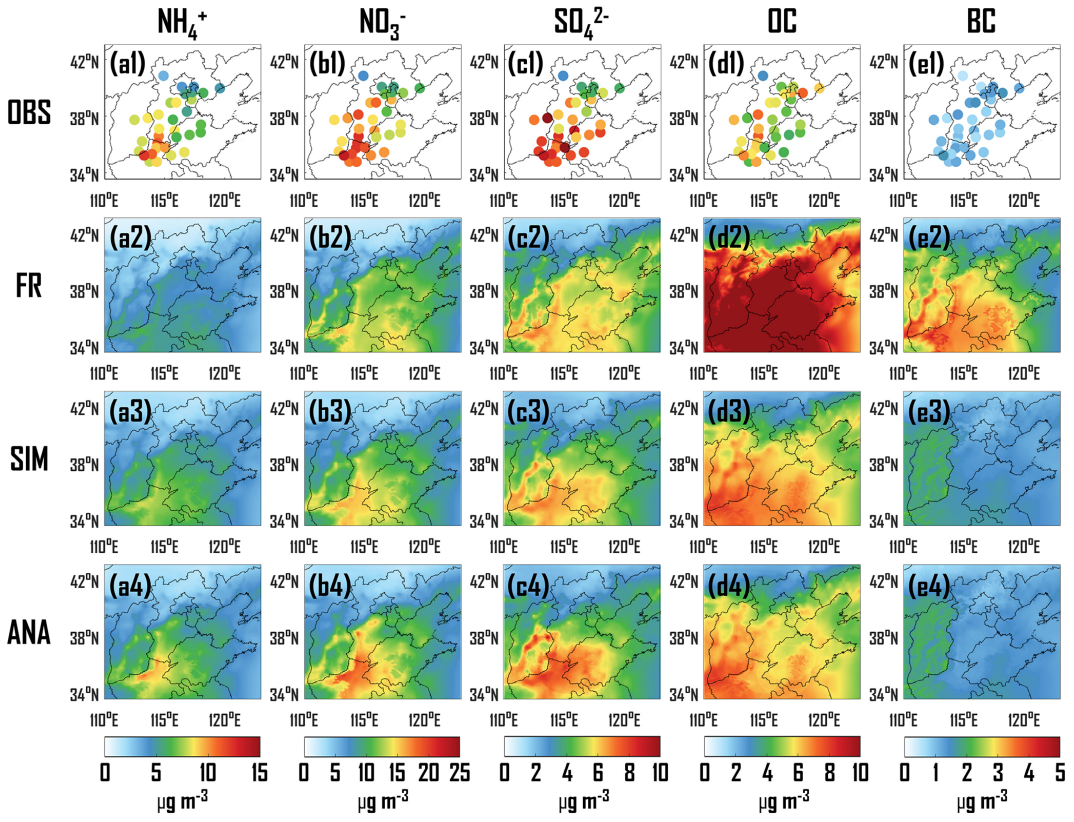

Figure 6. Spatial distribution of observation (OBS), free-run field (FRFR), ML-simulated background field (SIM) and analysis field (ANA) for NH_4^+ **(a1–a4)**, NO_3^- **(b1–b4)**, SO_4^{2-} **(c1–c4)**, OC **(d1–d4)** and BC **(e1–e4)**.

dressing the underestimation of NH_4^+ concentrations in the whole domain.

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 and c1). Although FR can capture the spatial patterns of NO_3^- and SO_4^{2-}, it significantly underestimates their concentrations (Fig. 6b2 and c2). Specifically, 63 %–79 % of DA sites and 89 % of VE sites underestimate by 1.87–3.76 and 1.57–3.44 $\mu g\,m^{-3}$, respectively (Fig. 7b1 and c1). Compared to FR, SIM mitigates the underestimations in the Hebei-Henan-Shandong junction areas and overestimations in the Beijing-Tianjin-Hebei eastern areas (Fig. 6b3 and c3), with improvements at most DA and VE sites (Fig. 7b2 and c2). After DA, ANA accurately characterizes the spatial distribution of NO_3^- and SO_4^{2-} concentrations (Fig. 6b4 and c4), with 88 %–100 % of DA sites and 56 %–67 % of VE sites merely underestimating by 0.77–1.31 and 1.85–2.73 $\mu g\,m^{-3}$, respectively (Fig. 7b3 and 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 and 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 and e1). However, FR significantly overestimated the concentrations of OC and BC in the North China region (Figs. 6d2, e2, and 7d1, e1), with an average overestimation of 6.12 $\mu g\,m^{-3}$ for OC and 1.99 $\mu g\,m^{-3}$ for BC at all DA sites, and 6.88 $\mu g\,m^{-3}$ for OC and 2.29 $\mu g\,m^{-3}$ for BC at all VE sites. Following incremental learning, SIM significantly reduced the overestimations (Figs. 6d3, e3, and 7d2, e2), resulting in an average overestimation of 1.46 $\mu g\,m^{-3}$ for OC and 0.53 $\mu g\,m^{-3}$ for BC at 71 %–79 % of DA sites, and 1.56 $\mu g\,m^{-3}$ for OC and 0.65 $\mu g\,m^{-3}$ for BC at 89 % of VE sites. The number of sites exhibiting overestimation and the degree of overestimation are markedly lower than those of FR. After DA, ANA further mitigates the overestimation in SIM, accurately interpreting the spatial distributions of OC and BC concentrations (Fig. 6d4 and e4), with the gaps between the observations and analysis fields for both DA and VE sites approaching 0 (Fig. 7d3 and e3). Assimilating the observations from 24 DA sites effectively mitigates the overestimation in the southern North China region (Fig. 7d4 and e4).

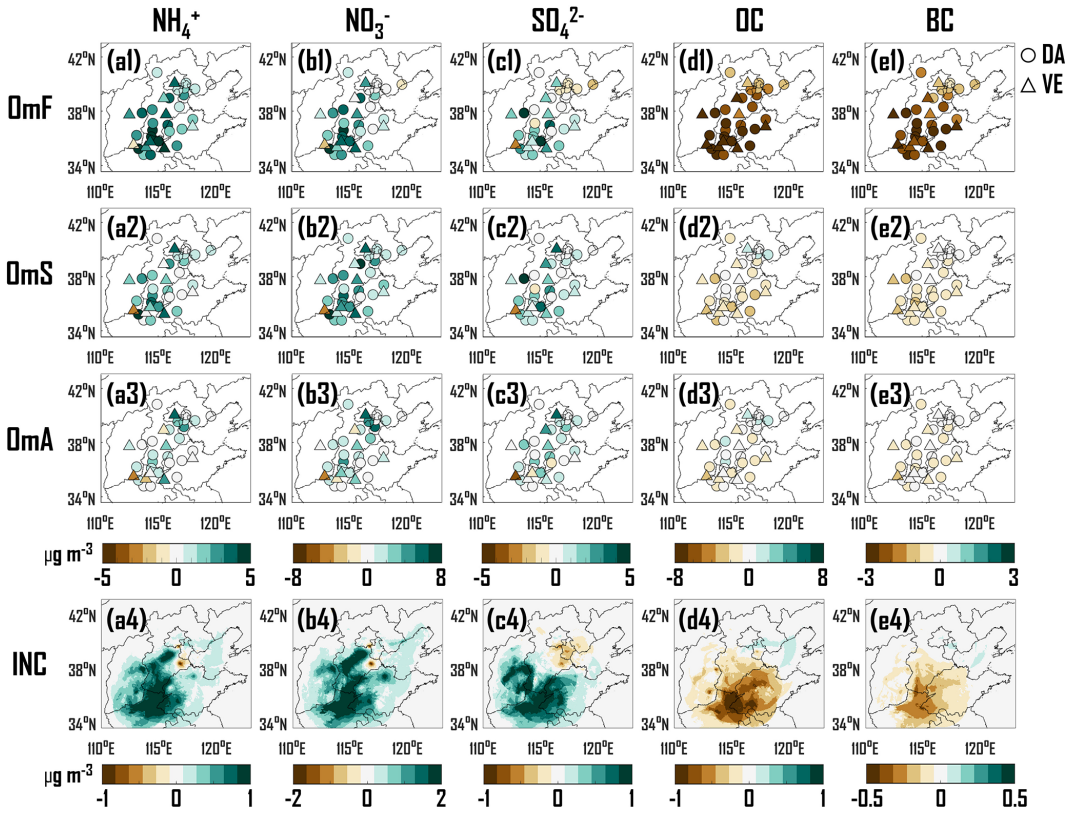

Figure 7. Spatial distribution of observation minus free-run field (OmF), observation minus ML-simulated background field (OmS), observation minus analysis field (OmA) and analysis field minus background field (INC) for NH_4^+ **(a1–a4)**, NO_3^- **(b1–b4)**, SO_4^{2-} **(c1–c4)**, OC **(d1–d4)** and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

3.3.3 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 ($0.25\,\mu g\,m^{-3}$) is significantly lower than that of NP2 ($0.81\,\mu g\,m^{-3}$), CAQRA ($1.18\,\mu g\,m^{-3}$), TAP ($0.92\,\mu g\,m^{-3}$), and Global-RA ($2.92\,\mu g\,m^{-3}$). Furthermore, the OmA of OIRF-LEnKF v1.0 is within $\pm1\,\mu g\,m^{-3}$ 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 ($0.19\,\mu g\,m^{-3}$) is significantly lower than that of NP2 ($0.93\,\mu g\,m^{-3}$), CAQRA ($8.42\,\mu g\,m^{-3}$), TAP ($2.24\,\mu g\,m^{-3}$), and Global-RA ($2.27\,\mu g\,m^{-3}$). Furthermore, the OmA of OIRF-LEnKF v1.0 is within $\pm2\,\mu g\,m^{-3}$ 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).

For SO_4^{2-} (Fig. 8c1–c5), the average absolute OmA of OIRF-LEnKF v1.0 ($0.54\,\mu g\,m^{-3}$) is slightly lower than that of NP2 ($0.86\,\mu g\,m^{-3}$) but significantly lower than that of CAQRA ($1.26\,\mu g\,m^{-3}$), TAP ($1.72\,\mu g\,m^{-3}$), and Global-RA

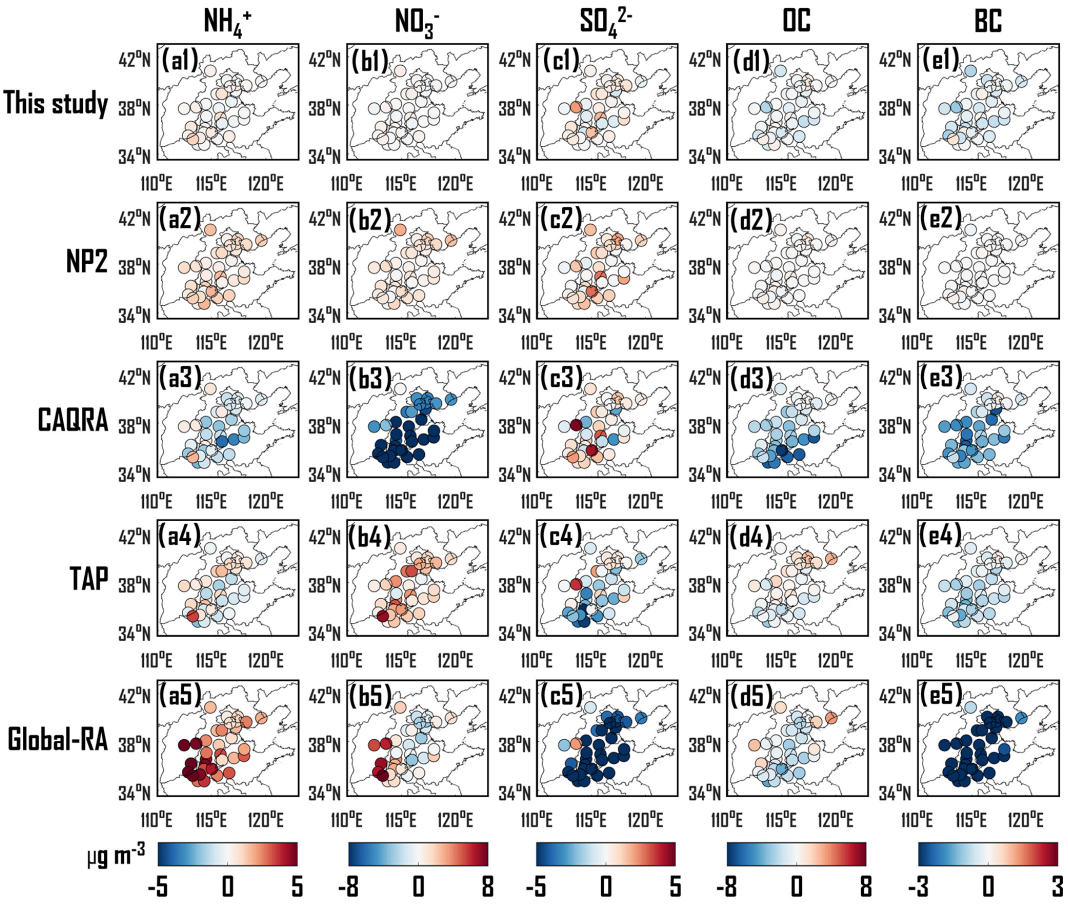

Figure 8. Difference between observations at a total of 33 sites and five reanalysis datasets for NH_4^+ **(a1–a5)**, NO_3^- **(b1–b5)**, SO_4^{2-} **(c1–c5)**, OC **(d1–d5)** and BC **(e1–e5)**. Global-RA is the combination of CAMSRA and MERRA-2.

($7.19\,\mu g\,m^{-3}$). 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 simulation 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 ($0.66\,\mu g\,m^{-3}$ for OC and $0.40\,\mu g\,m^{-3}$ for BC) is slightly higher than that of NP2 ($0.23\,\mu g\,m^{-3}$ for OC and $0.03\,\mu g\,m^{-3}$ for BC) but significantly lower than those of CAQRA ($2.90\,\mu g\,m^{-3}$ for OC and $1.32\,\mu g\,m^{-3}$ for BC), TAP ($1.04\,\mu g\,m^{-3}$ for OC and $0.65\,\mu g\,m^{-3}$ for BC), and Global-RA ($1.62\,\mu g\,m^{-3}$ for OC and $5.85\,\mu g\,m^{-3}$ 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: $1.12\,\mu g\,m^{-3}$, Fig. 9g) are significantly lower than those of other datasets (mean RMSE: 2.55–$8.52\,\mu g\,m^{-3}$, 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 and 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: $1.49\,\mu g\,m^{-3}$, Fig. 9g) are slightly worse than those of NP2 (mean

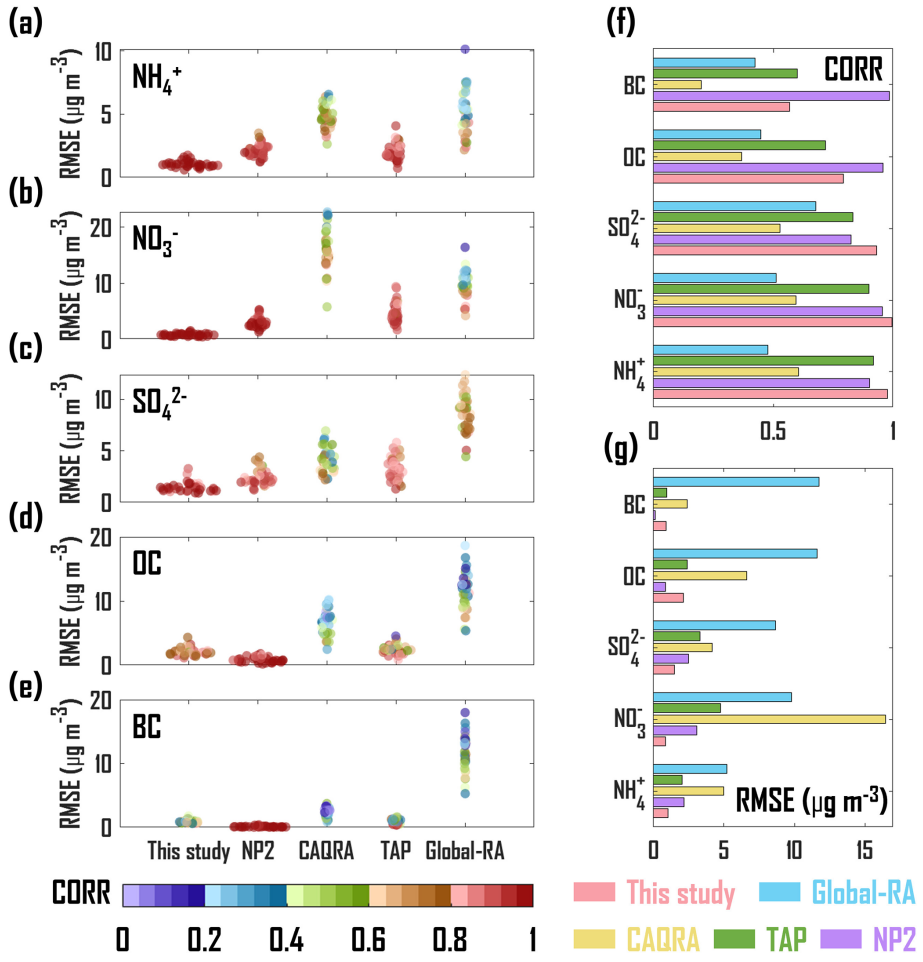

Figure 9. Pearson correlation coefficient (CORR) and root mean square error (RMSE, μg m^{-3}) quantified by the five reanalysis datasets and observations at a total of 33 sites for NH_4^+ **(a)**, NO_3^- **(b)**, SO_4^{2-} **(c)**, OC **(d)** and BC **(e)**. The averages of CORR **(f)** and RMSE **(g)** across all observational sites for the five reanalysis datasets for the five $PM_{2.5}$ chemical components. Global-RA is the combination of CAMSRA and MERRA-2.

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 model and the frequency of incremental learning, as well as by adopting hybrid nonlinear DA algorithms.

3.4 Limitations

Although the OIRF model serves as an efficient surrogate for the CTM in generating simulation or forecast ensembles for data assimilation, it inherits a constrained extrapolation capability of tree-based models. Specifically, the OIRF model may exhibit a tendency to saturate at learned extremes when extrapolating beyond its training data distribution, which directly limits its generalizability in diverse and complex atmospheric scenarios, such as the pollution extremes in seasons outside the training period. The poor performance of tree-based models on testing sets has been reported in our previous study (Li et al., 2025). Our incremental learning mechanism is designed to mitigate the extrapolation limitation by dynamically updating the RF model with new knowledge. However, the effectiveness of incremental learning is contingent upon the availability of high-quality analysis fields. A lack of observations, which prevents the generation of analysis fields, exposes the OIRF model to its inherent extrapolation limitations, leading to compromised simulation accuracy.

Replacing the RF model with an ensemble of deep neural networks (DNNs) holds promise for superior nonlinear mapping and extrapolation. However, the considerably higher

computational cost required for both training and inference of DNNs (Debjyoti and Utpal, 2025; Xi, 2022) results in an operational bottleneck that the process of updating and running an ensemble of DNNs can be slower than traditional CTM-based ensemble simulations, which could offset its accuracy advantages. Therefore, balancing the inherent predictive performance of a machine learning model against its computational cost remains a central challenge for the practical online coupling of machine learning with data assimilation.

4 Conclusions

In this paper, we online coupled the OIRF model with the LEnKF algorithm to develop a novel DA system (OIRF-LEnKF v1.0) that mitigates the limitations of high computational costs and inadequate advancements in generating background 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 an incremental learning 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, primarily because the simulation process requires only 0.13 % to 0.20 % of the CTM computation time, demonstrating its superior computational efficiency.

Sensitivity tests reveal that the background fields in OIRF-LEnKF v1.0 are more sensitive to updating frequency within the incremental learning mechanism. In contrast, the analysis fields exhibit a marked sensitivity to ensemble size. Specifically, the ΔCORR rises by 2.43 %–11.75 %, and the ΔRMSE decreases by 32.55 %–40.36 % when comparing a 1 h update frequency to the scenario without incremental learning during the simulation phase. Additionally, the ΔCORR increases by 9.75 %–19.04 %, and the ΔRMSE decreases by 16.70 %–30.48 % when comparing an ensemble size of 200 to that of 20 during the DA analysis phase. However, the 1 h update frequency diminishes the dependence of the analysis fields on ensemble size. Thus, an ensemble size of 50 with a 6 h update frequency is configured to balance computational efficiency, ML simulation 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 simulation 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 simulation capability of ML-based DA. Additionally, the spatial patterns of the background 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 $\mu g\,m^{-3}$) are markedly lower than those of the four reanalysis datasets (2.55–8.52 $\mu g\,m^{-3}$). 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, including observation data, modelling domain data, sensitivity test data and OIRF-LEnKF v1.0 output data, are openly accessible at https://doi.org/10.5281/zenodo.17346786 (Li and Yang, 2025). The open-access reanalysis datasets of CAQRA (https://doi.org/10.11922/sciencedb.00053, Tang et al., 2020; Kong et al., 2021), CAQRA-aerosol (https://doi.org/10.1007/s00376-024-4046-5, Kong et al., 2025) and ERA5 (https://doi.org/10.24381/cds.bd0915c6, Hersbach et al., 2023) from February to March 2022 were downloaded for the development and realization of OIRF-LEnKF v1.0 system, which have been packaged and uploaded at a repository (https://doi.org/10.5281/zenodo.17359290, Li, 2025) for easier access. The open-access reanalysis datasets of TAP (http://tapdata.org.cn, last access: 2 June 2025, Liu et al., 2022), NP2 (https://doi.org/10.5281/zenodo.10886914, Li et al., 2024b), CAMS (https://doi.org/10.24381/d58bbf47, Copernicus Atmosphere Monitoring Service, 2020; Inness et al., 2019) and MERRA-2 (https://disc.gsfc.nasa.gov/datasets?project=MERRA-2, last access: 2 June 2025, Randles et al., 2017) during February 2022 were downloaded for the evaluation of OIRF-LEnKF v1.0 sys-

tem, which have been packaged and uploaded at a repository (https://doi.org/10.5281/zenodo.17359290, Li, 2025) for easier access.

Supplement. The supplement related to this article is available online at https://doi.org/10.5194/gmd-19-4835-2026-supplement.

Author contributions. HL implemented the data assimilation system, performed the numerical experiments, conducted the analysis, and wrote the paper. TY conceived and designed the overall research framework, provided scientific guidance, wrote the paper, and devised the strategy for the responses and manuscript revision. LK and XT provided help for the system code and the CAQRA reanalysis dataset. DZ and GT provided $PM_{2.5}$ chemical component data. ZW did overall supervision. All authors reviewed and revised this paper.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. The authors bear the ultimate responsibility for providing appropriate place names. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

Acknowledgements. We thank for the technical support of the National Large Scientific and Technological Infrastructure "Earth System Numerical Simulation Facility" (https://cstr.cn/31134.02.EL, last access: 20 December 2025), 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 research has been supported by the National Natural Science Foundation of China (grant nos. 42422506 and 42275122).

Review statement. This paper was edited by Klaus Klingmüller and reviewed by two anonymous referees.

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. Imag., 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.

Copernicus Atmosphere Monitoring Service: CAMS global reanalysis (EAC4), Copernicus Atmosphere Monitoring Service (CAMS) Atmosphere Data Store [data set], https://doi.org/10.24381/d58bbf47, 2020.

Debjyoti, G. and Utpal, R.: Comprehensive Benchmark Study of Machine Learning and Deep Learning Approaches for Human Activity Recognition using the UCI HAR Dataset, Int. J. Comput. Appl., 187, 66–69, https://doi.org/10.5120/ijca2025925797, 2025.

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 Dynam., 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 T. Math. Softw., 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?, Philos. T. Roy. Soc. A, 379, 20200089, https://doi.org/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://doi.org/10.5555/3020751.3020778, 2014.

Gohari, K., Sheidaei, A., Yitshak-Sade, M., Colicino, E., and Kloog, I.: Exploring multivariate machine learning frameworks to parallelize $PM_{2.5}$ simultaneous estimations across the continental United States. Environ. Pollut., 374, 126161, https://doi.org/10.1016/j.envpol.2025.126161, 2025.

Gottwald, G. A. and Reich, S.: Supervised learning from noisy observations: Combining machine-learning techniques with data assimilation, Physica D, 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.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., and Thépaut, J.-N.: ERA5 hourly data on pressure levels from 940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS) [data set], https://doi.org/10.24381/cds.bd0915c6, 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 Syst., 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, hhttps://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.

Jalali, M. W., Saidi, B., Farahmand, H., Panah, M. A. R., and Saruhan, E. N.: Scalable AI-driven air quality forecasting and classification for public health applications, Discov. Atmos., 3, 25, https://doi.org/10.1007/s44292-025-00052-8, 2025.

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., 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.

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.

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), 14–16 August 2024, Athens, Greece, 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. Roy. 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 Syst., 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.: OIRF-LEnKF v1.0 related open-access datasets, Zenodo [data set], https://doi.org/10.5281/zenodo.17359290, 2025.

Li, H. and Yang, T.: OIRF-LEnKF v1.0, Zenodo [code and data set], https://doi.org/10.5281/zenodo.17346786, 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, 2024a.

Li, H., Yang, T., and Wang, H.: NAQPMS-PDAF v2.0 (Version 2.0), Zenodo [data set], https://doi.org/10.5281/zenodo.10886914, 2024b.

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, 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, 3, 100208, https://doi.org/10.1016/j.xinn.2022.100208, 2022.

Lin, G. Y., Chen, H. W., Chen, B. J., and Chen, S. C. et al.: A machine learning model for predicting $PM_{2.5}$ and nitrate concentrations based on long-term water-soluble inorganic salts datasets at a road site station, Chemosphere, 289, https://doi.org/10.1016/j.chemosphere.2021.133123, 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, 787–793, https://doi.org/10.5220/0007555207870793, 2019.

Liu, K., Zhang, Y., He, H., Xiao, H., Wang, S., Zhang, Y., Li, H., and Qian, X.: Time series prediction of the chemical components of $PM_{2.5}$ based on a deep learning model, Chemosphere, 342, 140153, https://doi.org/10.1016/j.chemosphere.2023.140153, 2023.

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.

Lv, L., Wei, P., Li, J., and Hu, J.: Application of machine learning algorithms to improve numerical simulation prediction of $PM_{2.5}$ and chemical components, Atmos. Pollut. Res., 12, 101211, https://doi.org/10.1016/j.apr.2021.101211, 2021.

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.-Atmos., 111, https://doi.org/10.1029/2005jd006149, 2006.

Meng, X., Hand, J. L., Schichtel, B. A., and Liu, Y.: Space-time trends of $PM_{2.5}$ constituents in the conterminous United States estimated by a machine learning approach, 2005–2015, Environ. Int., 121, 1137–1147, https://doi.org/10.1016/j.envint.2018.10.029, 2018.

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. Roy. 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 Min. 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. Climate, 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., and 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. Intel. 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, Proc. 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.: Multi-faceted 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.-Atmos., 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.

Tang, X., Kong, L., 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.: High-resolution Air Quality Reanalysis Dataset over China (CAQRA), Science Data Bank [data set], https://doi.org/10.11922/sciencedb.00053, 2020.

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. Product., 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., and Chen, X.: ChinaHighPMC: Daily Seamless 1 km Ground-Level $PM_{2.5}$ Composition Dataset for China (2000–Present) [Data set]. In Environmental Science & Technology (Version 1, Vol. 57, Issue 46, pp. 18282–18295), Zenodo [data set], https://doi.org/10.5281/zenodo.10011898, 2022.

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, 2019.

Xi, E.: Image Classification and Recognition Based on Deep Learning and Random Forest Algorithm, Wirel. Commun. Mob. Com., 2013181, https://doi.org/10.1155/2022/2013181, 2022.

Xie, T., Wang, C., and Peng, Y.: hi-RF: Incremental Learning Random Forest for Large-Scale Multi-class Data Classification, 2016/11, Atlantis Press, 312–321, https://doi.org/10.2991/aiie-16.2016.72, 2016.

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. Lett., 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.-Atmos., 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.