Authors' responses to Referees' comments

Journal: Geoscientific Model Development

Manuscript Number: egusphere-2025-3960

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

Authors: Hongyi Li, Ting Yang, et al.

Note:

Comment (12-point black italicized font).

Reply (indented, 12-point blue normal font).

"Revised text as it appears in the text (in quotes, 12-point blue italicized font)".

Anonymous Referee #1

1 General Comments:

This paper presents a method called the optimized incremental random forest ensemble forecasting model with the localized ensemble Kalman filter (OIRF-LEnKF), which combines an computationally lightweight emulator of $PM_{2.5}$ chemical components with the efficient and accurate data assimilation method LEnKF. This paper also presents a mechanism for online learning to update the OIRF model as new data arrives. The results on a real-world assimilation task of $PM_{2.5}$ concentrations in a region in China show that the proposed algorithm is effective in remaining stable across a long assimilation horizon while effectively assimilating observations that lead the analyses to remain close to a notion of ground truth (based on reanalysis). ***The use of machine learning in various data assimilation applications is an important investigation****, however, the authors could provide more motivation for particular choices made while creating the OIRF-LEnKF algorithm and perhaps further contextualize this work in relation to related work.*

Authors' response:

We sincerely thank the reviewer for the thoughtful review of our manuscript. We fully agree that a more robust justification for our specific methodological choices, as well as further contextualization of our work within the existing work, would enhance

the quality of our research.

In our detailed point-by-point responses below, we have addressed all scientific comments. Accordingly, in the revised manuscript, we have enhanced methodological motivation and contextualized this work in relation to related work.

2 Scientific Comments:

1) The authors propose a random forest model that is incrementally optimized as new information about the system is obtained, but provides minimal motivation for the choice of random forest over other approaches. This application area involves spatial datasets, which make a neural network architecture like a CNN a good fit, especially given that many state-of-the-art emulators of spatial systems rely in part on CNNs. The motivation for the choice of RF should be made more clear in the paper, and perhaps an ablation against a CNN-type architecture should be provided. To create an ensemble like a random forest, the CNN training could be bagged.

Authors' response:

We thank the reviewer for this insightful and constructive comment. We agree that providing stronger motivation for our choice of the Random Forest (RF) model is crucial, and we appreciate the suggestion to consider deep neural networks (such as CNNs). In the revised manuscript, we have expanded our justification for selecting RF as the surrogate model for atmospheric chemistry transport modeling, based on the following key considerations.

a. Computational Feasibility for Incremental Learning. Our proposed machine learning (ML) -based data assimilation (DA) framework requires the frequent retraining and updating of ML model members using newly assimilated analysis fields. However, an ensemble CNN approach is less optimal for this incremental learning mechanism. As presented in **Table R1**, under comparable hardware conditions, the training cost for a single CNN model is approximately an order of magnitude higher than that for an RF model (Xi, 2022; Debjyoti and Utpal, 2025). Consequently, serially constructing a 50-member ensemble CNN would incur a training time cost roughly 500 times greater than

that of an RF model. While CNN can indeed be bagged to create an ensemble, this approach is computationally prohibitive for the frequent ML model updates required by ML-DA framework.

b. Balanced Performance and Inference Speed for Prediction of $PM_{2.5}$ and its chemical components. Recent studies indicate that the predictive accuracy of RF for $PM_{2.5}$ and its chemical components is comparable to that of advanced deep learning models (Abuouelezz et al., 2025; Chen et al., 2023; Li et al., 2025). Notably, under the same hardware constraints, the inference speed of an RF model can be approximately 10 times faster than that of comparable CNN or Transformer models (Jalali et al., 2025). Consequently, RF models can achieve forecasting accuracy comparable to CNN models at a fraction of the computational cost, which is essential for maintaining the timeliness of the iterative forecast-assimilation cycle in our proposed system.

Introduction, Line 90-97: "*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 computational 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.*"

Table R1. Literature review of performance comparison between RF and CNN

Task type	Training cost		Inference speed		Error rate		Citation
	RF	CNNs	RF	CNNs	RF	CNNs	
Classification	20 minutes	3 hours	\	\	10.02%	9.67%	Xi, 2022
Recognition	12.7s per fold	186.4s per fold	0.8s per fold	0.2s per fold	2.90%	2.09%	Debjyoti and Utpal, 2025
Classification	\	\	0.028s per sample	0.218s per sample	\	\	Jalali et al., 2025
Regression	\	\	\	\	0.73-2.43% for training set	7.85-20.05% for training set	Li et al., 2025
Regression	\	\	\	\	17.74%	18.06%	Chen et al., 2023
Regression	\	\	\	\	0.216 for 1-h $PM_{2.5}$ forecast	0.213 for 1-h $PM_{2.5}$ forecast	Abuouelezz et al., 2025

Reference

Abuouelezz, W., et al.: Exploring $PM_{2.5}$ and PM_{10} ML forecasting models: a comparative study in the UAE. Sci Rep, 15, 9797, https://doi.org/10.1038/s41598-025-94013-1, 2025.

Chen, M.-H., et al.: $PM_{2.5}$ Concentration Prediction Model: A CNN-RF Ensemble Framework, Int. J. Environ. Res. Public Health, 20, 4077. https://doi.org/10.3390/ijerph20054077, 2023.

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

Jalali, M.W., et al.: 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.

Li, H., et al.: 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.

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.

2) Have there been other successful ML models of $PM_{2.5}$ chemical components? If so, they should be cited in the related work. If these emulators do exist, why were any of them not used instead of the authors' proposed random forest model?

Authors' response:

We thank the reviewer for raising this important point regarding related work and model selection.

Q1: Have there been other successful ML models of $PM_{2.5}$ chemical components?

A1: **Table R2** summarizes the successful ML models for the prediction of $PM_{2.5}$ chemical components. As replied to the Scientific Comment #1, the related work has been cited in the ***Introduction*** of the revised manuscript.

Table R2. Literature review of ML models used for the prediction of $PM_{2.5}$ chemical components

ML models	Target features	Performance on testing set	Citation
RF	Sulfate, nitrate, organic carbon, elemental carbon	R^2: 0.71-0.86	Meng, et al., 2018
RF	Sulfate, nitrate, ammonium, organic carbon, elemental carbon	R: 0.71-0.81	Lv et al., 2021
RF	Nitrate	R^2: 0.58	Lin et al., 2022
CNN-LSTM	Sulfate, nitrate, organic carbon, elemental carbon, crustal metals	R^2: 0.87-0.96	Liu et al., 2023
4D-STDF	Sulfate, nitrate, ammonium, chloride	CV-R^2: 0.66-0.75	Wei et al., 2023
CNN-BiLSTM	Sulfate, nitrate, ammonium, organic carbon, elemental carbon	R^2: 0.81-0.97	Li et al., 2025
PLS-SVM	Sulfate, nitrate, ammonium, organic carbon, elemental carbon, crustal metals	R^2: 0.82-0.98	Khuzestani et al., 2025
RF	Calcium, elemental carbon, silicon, sulfate	Spatial R^2: 0.93-0.95	Gohari et al., 2025

Q2: Why were any of them not used instead of the authors' proposed random forest model?

A2: As replied to the Scientific Comment #1, the RF model was selected as an optimal compromise between computational efficiency and predictive accuracy for the ensemble framework. Although **Table R2** indicates that RF's extrapolation capability

may limit its performance on testing set compared to other deep neural networks, we have designed an incremental learning mechanism to allow the RF model to continually adapt to new data distributions. However, the incremental learning mechanism relies on the availability of analysis fields assimilated observations. During periods of missing observations, the RF model remains susceptible to its poor extrapolation capability. We fully agree that ensemble CNNs could be more effective than RF in capturing nonlinear relationships, particularly for spatially structured data. In response to the reviewer's concern, we have added a dedicated discussion on the limitations of RF (now *Section 3.4*) and have renumbered the original *Section 3.4* as *Section 3.3.3*.

Section 3.4, Line 621-637: "***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."

Reference

Gohari, K., et al.: 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.

Khuzestani, R.B., et al.: Advancing Particulate Matter Chemical Composition Analysis: A Hybrid Machine-Learning Approach with UV-Vis Spectroscopy. Aerosol Sci. Eng., https://doi.org/10.1007/s41810-025-00282-8, 2025.

Li, H., et al.: 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.

Lin, G. Y., 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.

Liu, K., et al.: 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

Lv, L., et al.: 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.

Meng, X., et al.: 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.

Wei, J., et al.: 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, 46, 18282-18295, https://doi.org/10.1021/acs.est.3c00272, 2023.

3) In line 90, a claim is made that increasing the number of ensemble members in the forecast "mitigates the underestimation of forecast error covariance". It certainly helps mitigate, but it is not an assured cure. The authors should modify the language to something like "helps mitigate" to make the statement more accurate.

Authors' response:

We thank the reviewer for this precise comment. We agree that the original wording was too absolute. We have modified the text as suggested.

Introduction, Line 101-102: "*…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…*"

4) Is the idea of throwing away decision trees that do not perform as well as a predefined threshold on the updated dataset a novel contribution of this paper, or has this approach been used elsewhere? If it has been used elsewhere, the previous works should be cited.

Authors' response:

We thank the reviewer for this insightful question regarding the novelty of our proposed Optimized Incremental Random Forest (OIRF). We are aware of related work on incremental RF, such as hi-RF (Xie et al., 2016). The hi-RF model discards decision trees (DTs) with high errors based on an out-of-bag (OOB) error threshold and retrains new DTs using a combined dataset of old and new data, which shares a conceptual similarity with our OIRF model. However, key methodological distinctions exist.

a. The hi-RF model constructs new DTs from a merged bootstrap sample of old and new data without hyperparameter tuning, which may result in the performance of the new DTs being inferior to that of the discarded DTs, introducing uncertainties into the incremental learning. In contrast, the OIRF model updates DT members within a parallel Bayesian optimization framework to optimize the new RF structure, mitigating the uncertainties in incremental learning.

b. The number of high-error DTs replaced in hi-RF is variable over time, since it

depends on a dynamic OOB error threshold. Large and variable replacements of DTs could introduce instability or drift into the estimation of background error covariance within the data assimilation cycle. Our OIRF method employs a threshold depending on a percentile of statistical errors, which allows for controlled and stable replacement of DT members. This design makes OIRF more suitable for stable and long-term online coupling with a data assimilation framework.

Following the reviewer's suggestion, we have cited the relevant work in the revised manuscript and have clarified the specific advancements of our proposed method.

Section 2.1.2, Line 162-168: "*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).*"

Section 2.1.2, Line 175-180: "*The incremental learning mechanism introduces a threshold (τ_p) to screen out the DTs with poor simulation performance. The threshold is defined as the p^{th} 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.*"

Section 2.1.2, Line 192-196: "*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."

Reference

Xie, T., et al.: hi-RF: Incremental Learning Random Forest for Large-Scale Multi-class Data Classification, Proceedings of the 2016 2nd International Conference on Artificial Intelligence and Industrial Engineering (AIIE 2016), https://doi.org/10.2991/aiie-16.2016.72, 2016.

3 Technical Corrections:

1) It would be helpful to spell out the name of the OIRF-LEnKF in the abstract (line 15).

Authors' response:

We thank the reviewer for this helpful suggestion. In the revised manuscript, we have replaced "OIRF-LEnKF" with its full name.

Abstract, Line 15-18: "…*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*…"

2) In line 55, "DA technique has been widely used [...]" should be corrected to either "DA has been widely used [...]" or "DA techniques have been widely used [...]."

Authors' response:

We thank the reviewer for catching this expression inaccuracy. We have removed the work "technique" in the revised manuscript.

Introduction, Line 57-58: "*…DA has been widely used to generate reanalysis datasets of $PM_{2.5}$ chemical components at global and national scales…*"

3) "Where" immediately after equation (1) should be lowercase.

Authors' response:

We thank the reviewer for pointing out this oversight. We agree and have corrected the capitalization accordingly in the revised manuscript.

Section 2.1.2, Line 156: "*where x represents the input features…*"

4) What criteria is used to determine a split for each decision tree? Is it based on MAE? This should be made more clear in the text (roughly around line 145).

Authors' response:

We thank the reviewer for this insightful question, which allows us to clarify an important methodological detail. The criterion for selecting the optimal split at each node during the training of an individual decision tree (DT) is the maximization of the reduction in Mean Squared Error (MSE). We have revised the manuscript to explicitly state that the splitting criterion for each DT.

Section 2.1.2, Line 159-160: "*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.*"

5) y is used to describe an analysis in line 154 but is then used to describe observations in line 215. The authors should stay consistent in the text that y refers to observations.

Authors' response:

We sincerely thank the reviewer for catching this inconsistency in our notation. In response to the Reviewer's suggestions, we revised the manuscript to stay consistent in the text that "y" refers to observations.

Section 2.1.2, Line 169-190: "

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

Here, f_n^{score} is the MAE value of the n^{th} DT. K is the total number of grids of $PM_{2.5}$ chemical component concentrations. x_i^{ana} is the analysis value of concentrations at the i^{th} grid point after DA. $f^{DT}(x_i, \theta_n)$ denotes the simulation value of the n^{th} DT at the i^{th} 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.

$$f_t^{DT} = \begin{cases} f^{DT}(x, \theta_{\mathrm{n}}|x_{t-\Delta t}^{ana}), f_{\mathrm{n}}^{score} \leq \tau_{\mathrm{p}}, \mathrm{n} = 1, 2, \ldots, \mathrm{N_p} \\ f^{DT}(x, \theta_{\mathrm{n}}|x_t^{ana}), f_{\mathrm{n}}^{score} > \tau_{\mathrm{p}}, \mathrm{n} = \mathrm{N_p} + 1, \mathrm{N_p} + 2, \ldots, \mathrm{N} \end{cases} , \quad (3)$$

Here, f_t^{DT} represents the final output of the updated DTs following incremental learning at time t. $f^{DT}(x, \theta_n|x_{t-\Delta t}^{ana})$ denotes the output of the retained old DTs while $f^{DT}(x, \theta_n|x_t^{ana})$ refers to the output of the new DTs. Δt represents the time interval of incremental learning. τ_p indicates the p^{th} percentile value of $f_n^{score}(n = 1, 2, \ldots, N)$, 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 ($f^{OIRF}(x)$) of the OIRF model at time t is derived from Eq. (4) by averaging the outputs of the updated DTs.

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

6) In lines 154-155, should "nth grid point after DA" be changed to the "ith grid point"? And similarly should "nth DT at the nth grid point" be changed to "nth DT at the ith grid point"?

Authors' response:

We thank the reviewer for this precise and correct observation. The revised version

can be found in the reply to **Technical Correction #5**.

7) Immediately after equation (5), it should be made clear that f_t^DT(x,theta_n) with a bar over the expression refers to the ensemble mean across decision trees in the random forest.

Authors' response:

We thank the reviewer for the suggestion. The revised version is as follows.

Section 2.1.3, Line 228-229: "*Here,* $\boldsymbol{P}_t^f$ *is the flow-dependent background error covariance matrix of* $PM_{2.5}$ *chemical component concentrations at time* t, $\overline{f_t^{DT}(x,\theta_n)}$ *refers to the ensemble mean across decision trees in the random forest at time* t."

8) I think that the paper would benefit from a mathematical formulation of the difference between domain localization and observation localization (in the section in lines 219-243).

Authors' response:

We thank the reviewer for this insightful suggestion. The fundamental update form of the EnKF with domain localization is analogous to the global EnKF (as presented in original Eq. (9) of the original manuscript) but uses forecast fields and observations within a specific localization radius. In response to the Reviewer's suggestion, we have added the mathematical formulation of domain localization.

Section 2.1.3, Line 253-262: "*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).*

$$x_{n,i}^{ana} = f_i^{DT}(x,\theta_n) + \boldsymbol{K_\delta}\left(y_\delta^o + y'^o_{n,\delta} - H_\delta\left(f_i^{DT}(x,\theta_n)\right)\right), n = 1,2,\dots,N, \quad (11)$$

Here, $x_{n,i}^{ana}$ *is the analysis value at* i^{th} *grid point of the* n^{th} *ensemble member.* $f_i^{DT}(x,\theta_n)$ *is the background value at* i^{th} *grid point of the* n^{th} *ensemble member.* $\boldsymbol{K}_\delta$

is the local Kalman gain matrix computed from the ensemble covariance within the localization domain δ. y_{δ}^{o} *is the observation of* $PM_{2.5}$ *chemical components within the localization domain* δ *and* $y'^{o}_{n,\delta}$ *is the observation perturbation of the* n^{th} *ensemble member within the localization domain* δ. H_{δ} *is the linear observation operator within the localization domain* δ."

9) The construction of W in equation (11) is not immediately clear. What values do i and j range from? Why is this matrix forced to be diagonal? What is the dimensionality of W? The answers to these questions should be made more clear in the text.

Authors' response:

a. i ranges from one to the total number of analysis grids within the whole domain, j ranges from one to the total number of observation sites within a localization domain.

b. The observation error covariance matrix **R** is assumed to be diagonal in practice, implying that observation errors are spatially uncorrelated and the observations can be processed serially (Nerger, 2015; Valler et al., 2019). The distance-based weight matrix **W** is consequently constructed as a diagonal matrix, applying a distance-dependent weighting directly to the diagonal elements of observation error covariance matrix **R** to attenuate the influence of observations that are farther from the target analysis point.

c. **W** is an $n \times n$ matrix, where n denotes the number of effective observations within the localization domain.

In response to the Reviewer's suggestion, we have revised the text.

Section 2.1.3, Line 282-287: "*The distance-based weight matrix (*$\boldsymbol{W_i}$*) for the* i^{th} *localization domain is obtained using a Gaussian function:*

$$\boldsymbol{W_i} = diag\left(exp\left(\frac{-d(i,j)^2}{2L^2}\right)\right), j = 1, 2, \dots, N_{obs} \ , \qquad (13)$$

Here, $d(i,j)$ *is the Euclidean distance between center grid point of the* i^{th} *localization domain and observation point* j. L *is the decorrelation length.* N_{obs} *is the total number of effective observations within the* i^{th} *localization domain.* $\boldsymbol{W}$ *is constructed as a diagonal matrix (*$N_{obs} \times N_{obs}$*), applying a distance-dependent weighting directly*

to the diagonal elements of observation error covariance matrix $\boldsymbol{R}_t$."

Reference

Nerger, L.: On Serial Observation Processing in Localized Ensemble Kalman Filters. Mon. Wea. Rev., 143, 1554-1567, https://doi.org/10.1175/MWR-D-14-00182.1, 2015.

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.

10) In Table 1, could the authors please also list the dimensions of the analysis (# latitudes, # longitudes, # features)?

Authors' response:

We thank the reviewer for the suggestion. The revised Table 1 is as follows.

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°N -44.90°N, 108.07°E-127.01°E)
	Spatial resolution	5 km×5 km
	Temporal resolution	1 h
	Meteorological input feature	U-component wind, V-component wind, temperature, relative humidity and geopotential
	Anthropogenic input feature	$PM_{2.5}$, PM_{10}, SO_2, NO_2, CO and O_3
	Ensemble size	2, 5, 10, 15, 20, 30, 40, 50, 100, 200
	Update frequency	0, 18-h interval, 12-h interval, 6-h interval, 1-h interval
	Hyperparameter for tuning	Minimum number of leaf node observations, maximal number of decision splits, and number of predictors to select at random for each split
	Optimization iteration	30
	Data partition	Re-partition at every iteration
Data assimilation	State dimension	5, including SO_4^{2-}, NO_3^-, NH_4^+, OC and BC
	Latitudinal dimension	249 grids
	Longitudinal dimension	300 grids
	Algorithm	LEnKF
	Localization radius	200 km
	Decorrelation length	80 km

11) In Figure 3a, is the objective referenced from the Bayesian optimization? It may be more clear to reference an equation number in the caption. In Figure 3c, why is there a sudden decrease at an ensemble size of 30 in the OIRF-LEnKF/NP2 (%)?

Authors' response:

We thank the reviewer for the suggestion. The objective value in Fig. 3a is derived from the Bayesian optimization. We have supplemented the objective function in Bayesian optimization and referenced an equation number in the caption.

Section 2.1.2, Line 213-216: "$J(\theta) = ln(1 + \frac{1}{N}\sum_{i=1}^{N}\left(y_i^{pred}(\theta) - y_i^o\right)^2)$, (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. $y_i^{pred}(\theta)$ is the predicted value for the i^{th} sample, y_i^o is the observation value for the i^{th} sample."

Figure 3a, Caption: "*...(a) Variation in the minimum objective value throughout the Bayesian optimization process and time consumed by each iteration, determined by Eq. (5)...*"

We attribute the sudden decrease in the OIRF-LEnKF/NP2 ratio at an ensemble size of 30 to a significant increase in the computational time required by the NP2 forecast. This fluctuation is likely related to inherent uncertainties in the two-level parallel structure of NP2. In NP2, ensemble members are distributed across multiple computing nodes in a cluster, while the grid points for a single member are further parallelized across multiple CPUs within a node (Li et al., 2024). This structure can occasionally lead to communication congestion or latency between different MPI communicators (the communicators used in NP2 are presented in Fig. R1 in Wang et al. (2022)'s study), resulting in an augmentation of additional computational cost. In contrast, the OIRF-LEnKF is designed to avoid such uncertainties. The state variables, namely five $PM_{2.5}$ chemical components, are processed independently on five separate computational nodes (Fig. 2 in our manuscript), eliminating the need for inter-node communication. We have highlighted the advantages of the OIRF-LEnKF parallel

architecture as follows.

Section 2.1.3, Line 267-269: "…*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*…"

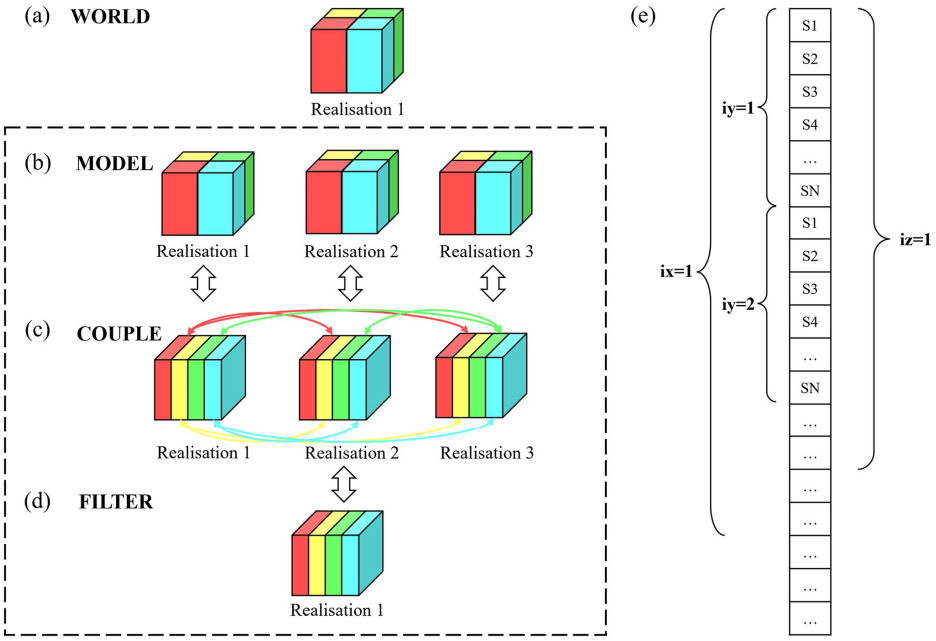

Fig. R1 in Wang et al. (2022)'s study

Reference

Li, H., Yang, T., Nerger, L., Zhang, D., Zhang, D., Tang, G., Wang, H., Sun, Y., Fu, P., Su, H., and Wang, Z.: NAQPMS-PDAF v2.0: a novel hybrid nonlinear data assimilation system for improved simulation of $PM_{2.5}$ chemical components, Geosci. Model Dev., 17, 8495-8519, https://doi.org/10.5194/gmd-17-8495-2024, 2024.

Wang, H., Yang, T., Wang, Z., Li, J., Chai, W., Tang, G., Kong, L., and Chen, X.: An aerosol vertical data assimilation system (NAQPMS-PDAF v1.0): development and application, Geosci. Model Dev., 15, 3555-3585, https://doi.org/10.5194/gmd-15-3555-2022, 2022.

12) The colorbar in all subfigures in Figure 4 should start at 0 so that perceived color variations more closely correspond to significant differences in the values in the table. Figure 4c, for example, has two different colors assigned to 0.77 in the bottom right corner, likely due to small differences past the third decimal place. If these heatmaps no longer look interesting after making this change, then another plotting technique highlighting any interesting aspects should replace Figure 4.

Authors' response:

We thank the reviewer for the suggestion. However, as the reviewer anticipated and we verified, setting the colorbar to start at 0 would compress the entire color spectrum into a very narrow range, since our performance metrics (CORR and RMSE) are all concentrated at a narrow scale (e.g., R from 0.70 to 0.84 at analysis step). This would make it impossible to discern the meaningful variations across the parameter space. To address this fundamental issue and fully adhere to the reviewer's suggestion of ensuring visual accuracy, we have chosen to display the percentage change of each metric relative to a defined performance baseline (e.g., minimum CORR and maximum RMSE) and we have utilized a more suitable colorbar.

Section 3.2, Line 412-425: "*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 2.42 % to 11.75 %, and the percentage change of RMSE relative to maximum RMSE (ΔRMSE) decreases by 32.55 % to 40.36 % when comparing a 1-hour update frequency to the scenario without incremental learning, which indicates that high-frequency incremental learning effectively enhances the adaptability of the statically trained ML model to the non-stationary data distributions, enabling it to demonstrate improved generalization capabilities and higher 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-hour 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 becomes more accurate, leading to improved ensemble simulation skill (Chen, 2024)."

Section 3.2, Line 438-440: "*…Specifically, the ΔCORR increased by 9.75 % to 19.04 %, and the ΔRMSE decreased by 16.70 % to 30.48 % when comparing an ensemble size of 200 to that of 20…*"

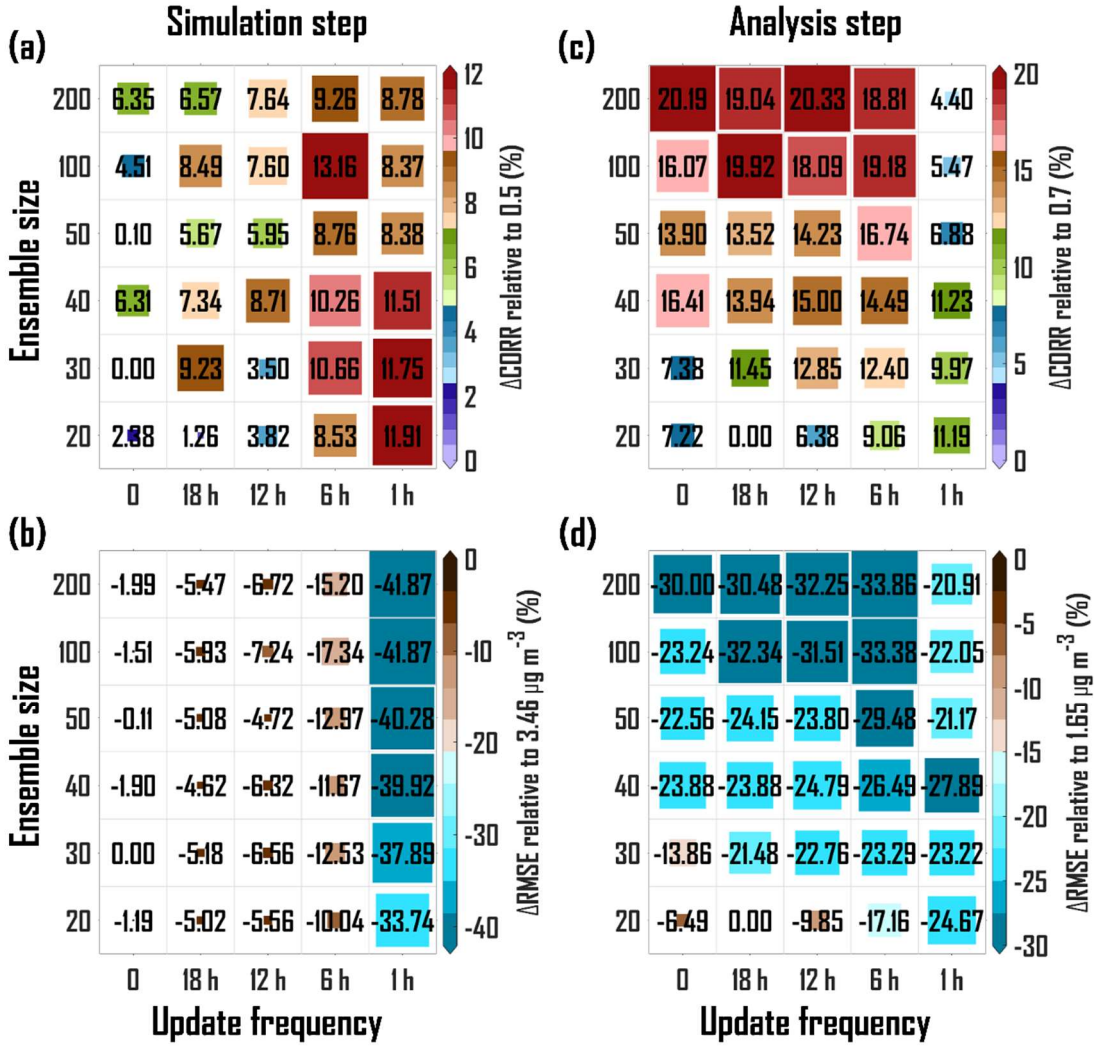

Figure 4*. (a) Percentage change of Pearson correlation coefficient (CORR) relative to the minimum CORR (0.5) (ΔCORR, %) for sensitivity test with six ensemble sizes (20, 30, 40, 50, 100, 200) and five update frequencies (no update, 18-hour interval, 12-hour*

interval, 6-hour interval and 1-hour 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.

Authors' responses to Referees' comments

Journal: Geoscientific Model Development

Manuscript Number: egusphere-2025-3960

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

Authors: Hongyi Li, Ting Yang, et al.

Note:

Comment (12-point black italicized font).

Reply (indented, 12-point blue normal font).

"Revised text as it appears in the text (in quotes, 12-point blue italicized font)".

Anonymous Referee #2

1 General Comments:

This manuscript presents "OIRF-LEnKF v1.0," a hybrid data assimilation (DA) system that couples an optimized incremental Random Forest (OIRF) model with a Localized Ensemble Kalman Filter (LEnKF). The primary goal is to address the computational inefficiency and limited forecasting improvement associated with traditional Chemical Transport Model (CTM)-based DA systems. By replacing the CTM ensemble with a machine learning (ML) ensemble that updates itself via incremental learning, the authors claim to achieve significant efficiency gains and improved accuracy in estimating $PM_{2.5}$ chemical components. The topic is highly relevant to the scope of Geoscientific Model Development, as it addresses the critical bottleneck of computational cost in atmospheric chemistry DA. The integration of incremental learning (updating decision trees based on analysis increments) is a novel and interesting approach to handling non-stationary error distributions. The validation against independent sites and other reanalysis datasets suggests the system performs well.

However, there are several major concerns regarding the terminology used (specifically "forecasting"), the dependence on reanalysis inputs, and the circularity of the self-evolving mechanism that must be addressed before publication. The critical distinction

between a "reanalysis generator" and a "forecast system" seems blurred in the current experimental design. The experimental period is very insufficient to support the claims.

Authors' response:

We sincerely thank the reviewer for the thorough and insightful review, as well as for the positive assessment of the novelty and relevance of our work. We appreciate the constructive criticisms, which have helped us identify crucial areas for clarification and improvement. We will address the major concerns regarding the terminology used (specifically "forecasting"), the dependence on reanalysis inputs, the circularity of the self-evolving mechanism, and the insufficient experimental period point-by-point below.

2 Major Comments:

1) Clarification of "Forecasting" vs. "Hindcasting/Reanalysis"

The title and abstract repeatedly emphasize the system's "forecasting" capability. However, Section 2.2.1 states that the input features for the OIRF model include meteorological parameters from ERA5 and atmospheric pollutants from CAQRA (a reanalysis dataset). In that setting, the model is effectively learning an instantaneous relationship (features at time t → components at time t). In a true operational forecast setting, ERA5 and CAQRA data are not available in real-time; they are retrospective datasets. If the OIRF model relies on con-current reanalysis data as inputs to predict chemical components, this is technically a "diagnostic" application, not a "prognostic forecast." The authors must clarify this distinction. If the system is intended for operational forecasting, they should discuss how it would perform using forecast meteorology (e.g., IFS or GFS) and forecast pollutants (e.g., CTM forecasts) as inputs, rather than high-quality reanalysis data. This is not a minor wording issue: if inputs are reanalysis fields at the verification time, then improvements in RMSE/CORR do not necessarily translate to operational forecast skill. If the primary purpose is generating reanalysis datasets (as implied by the comparison in Section 3.4), the manuscript should be reframed to reflect that this is a "reanalysis system" or "hindcast system,"

as calling it a "forecast" is misleading given the input data latency.

Authors' response:

We sincerely thank the reviewer for this crucial and insightful comment, which correctly identifies the most important interpretive limitation of our current model input design. We fully agree that the distinction between a diagnostic application and a real-time prognostic forecast is fundamental.

a. Terminology correction

The reviewer is correct. The OIRF model learns an instantaneous mapping relationship based on retrospective reanalysis datasets and technically performs a "reanalysis-based simulation" rather than a "prognostic forecast". Therefore, referring to simulation or instantaneous mapping as an operational "forecast" is misleading. We apologize for this imprecision. **In the revised manuscript, we have thoroughly replaced the inappropriate terms such as "forecast", "forecast field", and "FOR" with more accurate terms such as "simulation", "background field", and "SIM" in the title, abstract, main text, figures (including Fig. 1, Fig. 3, Fig. 4, Fig. 5, Fig. 6, Fig. 7) and tables.** For brevity, only the revised title and abstract are shown below, and all modified figures are displayed at the end of this reply.

Title, Line 1-4: "*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*"

Abstract, Line 12-29: "*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, particularly during the simulation process (0.13-0.20 %). Sensitivity tests demonstrate that the self-evolution mechanism in our system enhances the Pearson correlation coefficient (CORR) and reduces the RMSE during the simulation process by 2.28-11.75 % and 32.94-40.98 %, respectively, compared to the stationary training mechanism. A 2-month DA experiment reveals that the RMSE values of chemical components after DA are less than 7.80 μg m^{-3} and 2.36 μg m^{-3} during the 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 μg m^{-3}), the dataset generated by OIRF-LEnKF v1.0 (CORR: 0.97, RMSE: 1.12 μg m^{-3}) demonstrates higher data quality."

b. Clarification of our study's primary goal

The primary objective of this work is to propose and validate a novel framework that online couples an incrementally updatable AI-based surrogate model and an ensemble data assimilation algorithm, which enables the AI-based surrogate model and the data assimilation component to benefit from the dynamic information provided by the other at each iteration. During the concept proof stage, using optimal reanalysis inputs is deliberate to establish a valid representation of forecast/simulation uncertainty. Specifically, the OIRF-LEnKF utilizes the decision tree members in the OIRF model to estimate the background error covariance without input perturbations, which implicitly assumes that the forecast/simulation uncertainty mainly originates from the OIRF model's inherent incompleteness in learning the mapping relationship between input and target features. Therefore, using reanalysis data as input excludes the additional

uncertainty that would arise from imperfect forecast inputs.

c. Discussion of operational forecasting

The application of the OIRF-LEnKF system in operational forecasting is feasible, but its comprehensive validation would require a broader set of experiments, such as sensitivity tests on forecast ensemble generation. The in-depth investigation on the operational forecasting extends beyond the primary scope of this paper. We sincerely thank the reviewer's constructive suggestions, which provide significant inspiration. Our immediate future work will indeed prioritize these forecasting experiments, such as employing forecast data as input and assessing performance under different ensemble generation strategies (e.g. using perturbed meteorological forecast data alone, jointly perturbing meteorological and pollutant forecasts, or developing hybrid methods that integrate input perturbations with the intrinsic ensemble statistics from the decision tree members).

2) Circularity / information leakage risk in the incremental learning loop

A key design choice is that each decision tree is scored using MAE against the analysis field at the same time step, and trees are replaced by new trees trained on "analysis" targets. I have two concerns on this design. The first one is that the system is self-training on its own analysis. The system progressively trains on DA outputs (which incorporate the observations), not solely on an external reference dataset. This can lead to overly optimistic performance if not carefully controlled. The second one is that the scoring target is non-independent. The analysis field is itself a function of the forecast ensemble (through the forecast covariance used by EnKF). Even with localization, using the analysis as "ground truth" for selecting trees can create a feedback loop where the ensemble is optimized to match its own internally constructed target. At minimum, the paper should include a leakage-aware evaluation, for example, scoring the incremental learning using withheld stations (VE sites) not assimilated, or withheld time blocks, rather than analysis fields produced by assimilating the same network. Providing an ablation where incremental learning is driven by an external target versus DA-derived analysis, to quantify how much of the gain comes from DA self-training can

also be very valuable.

Authors' response:

We thank the Reviewer for this exceptionally insightful comment.

Regarding the first concern, we fully agree that the system progressively trains the OIRF model using DA outputs can lead to overly optimistic performance if not carefully controlled. We would like to clarify several key points as follows.

a) DA outputs serve as a high-quality target set for training, aiming to accurately establish the mapping relationship between the input features and five $PM_{2.5}$ chemical components. Importantly, the re-training utilizes the DA outputs from the previous time step. The updated RF model is then applied to provide background fields for the current and all subsequent time steps until the next update. This design ensures that the model does not gain prior knowledge of future states, thereby preventing artificially optimistic performance.

b) An external and independent target set of high-quality observations typically provides insufficient sample size (only 9 VE sites in our case) for training at a single time step. Meanwhile, as validated in *Section 3.4*, external reanalysis datasets exhibit lower accuracy than DA outputs, making them a suboptimal choice for the "ground truth" target in the re-training process.

c) **Most crucially, our system incorporates specific controls (update frequency and update intensity) to mitigate the risk of overly optimistic performance**. The update frequency parameter, which determines how often the OIRF model integrates DA outputs, is optimized by sensitivity experiments, as detailed in *Section 3.2*. The update intensity parameter is implemented by a controllable threshold (τ_p), which governs the proportion of decision trees (DTs) replaced. A higher threshold ensures that only a small and stable fraction of the DTs is replaced during each update cycle, which prevents model overfitting to the new DA outputs.

We acknowledge that some rationales and controls were not explicitly detailed in the original manuscript, which may have caused confusion. We have revised the text accordingly to provide the necessary clarification as follows.

Section 2.1.1, Line 128-143: "*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."

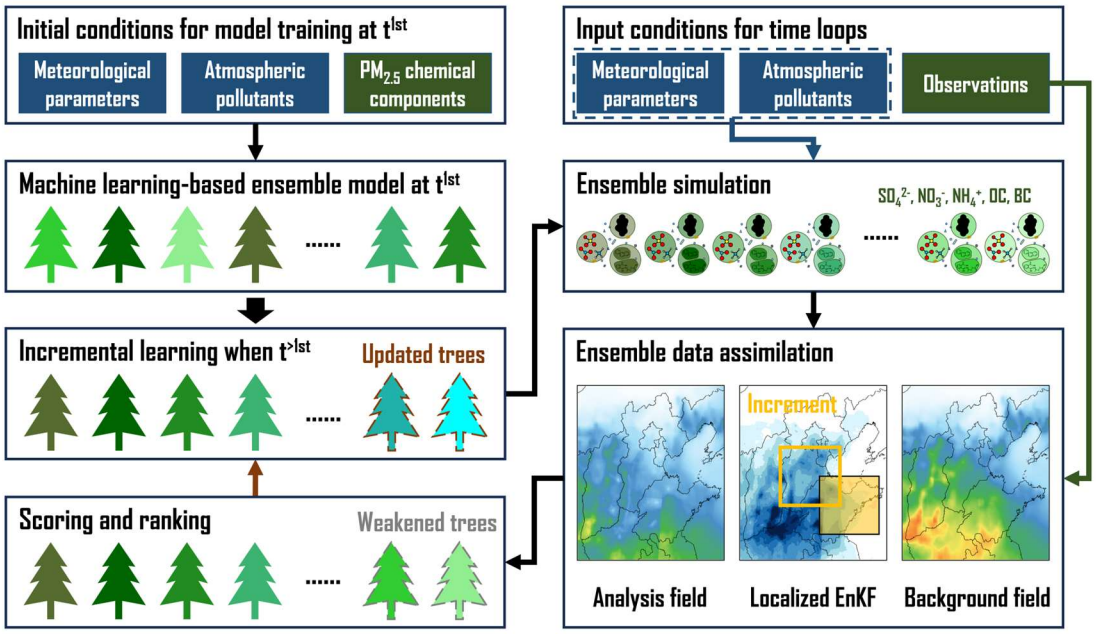

Figure 1*. The framework of OIRF-LEnKF v1.0.*

Section 2.1.2, Line 175-180: "*The incremental learning mechanism introduces a threshold (τ_p) to screen out the DTs with poor simulation performance. The threshold is defined as the p^{th} 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.*"

Section 2.1.2, Line 185-186: "*...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.*"

Regarding the second concern, we fully agree that our work should include a leakage-aware evaluation. Following the Reviewer's suggestion, we have conducted an experiment in which we score the incremental learning using the observations from VE sites that have not been assimilated.

Section 2.1.2, Line 162-168: "*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)."

Supplement: "***Sect. S1: Leakage-aware evaluation of incremental learning***

In the incremental learning mechanism, each decision tree (DT) member is scored by comparing its simulation to the analysis field using mean absolute error (MAE). However, using the analysis field as the scoring target for selecting trees could arise a feedback loop risk as the DT ensemble may become optimized toward its own internally constructed target. Therefore, we conducted a leakage-aware evaluation for February 2022 by comparing simulation performance of the OIRF model when the scoring target is set as the analysis field against when it is set as the independent observation at withheld sites (VE sites) not assimilated. Fig. S1 shows that both scoring targets achieved comparable performance across all five $PM_{2.5}$ chemical components, with correlation coefficient (CORR) values of 0.39-0.85 (analysis-field target) versus 0.39-0.86 (independent-observation target), and RMSE values of 1.02-5.85 μg m^{-3} (analysis-field target) versus 0.95-5.68 μg m^{-3} (independent-observation target). This finding suggests that the theoretical risk of a feedback loop from using the analysis field as the scoring target was limited during the study period. Adopting an independent-observation target is recommended in practice, since it yields slightly superior skill and fully eliminates the theoretical concern of an information leakage risk.

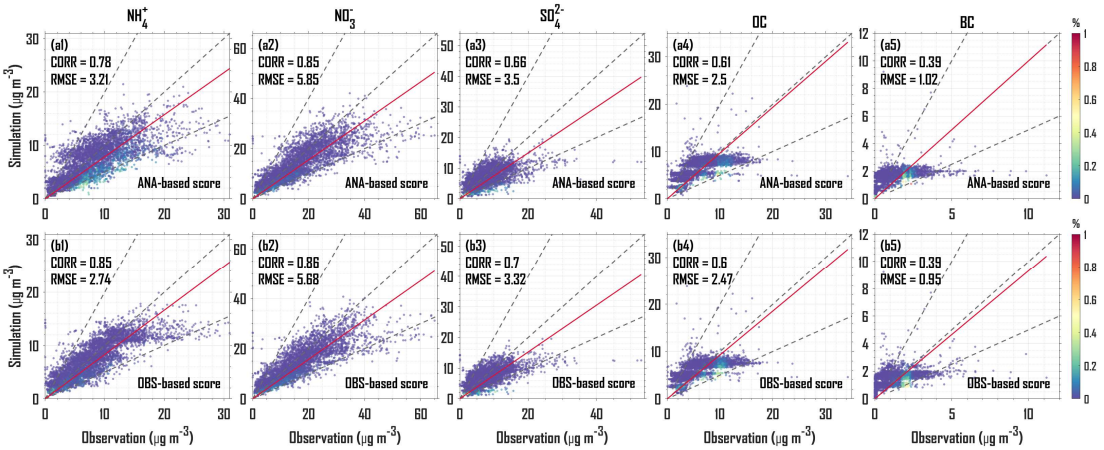

Figure S1: *Scatterplots with probability density of simulated versus observed mass*

concentrations at independent VE sites correspond to the two scoring targets used in the incremental learning process, including analysis fields (ANA) (a1-a5) and independent observations (OBS) (b1-b5). The gray dotted lines represent the 2:1, 1:1, and 1:2 lines, and the red solid line represents the fitting regression line."

3) Insufficient experimental period limiting model extrapolation and generalizability

The experimental validation is strictly limited to a two-month period (February–March 2022). This short duration fundamentally undermines the manuscript's claims regarding the system's robustness and its "self-evolving" capability, primarily due to the inherent limitations of the chosen machine learning architecture.

From the ML perspective, The OIRF model relies on the Random Forest (RF) algorithm. A well-known limitation of tree-based methods is their inability to extrapolate beyond the range of values encountered in the training data. By restricting the training and validation to a single two-month window, the model is only exposed to a specific subset of atmospheric conditions. If the system encounters pollution episodes more severe or chemically distinct than those in the February-March training set, the RF model will likely "clip" the forecast to the maximum value previously learned, failing to capture new extremes. The current experimental design does not demonstrate that the "incremental learning" mechanism can overcome this fundamental extrapolation barrier when faced with out-of-distribution data.

From the physics perspective, a system trained and validated exclusively on winter/early spring data cannot validly be claimed as a "Self-evolving Data Assimilation System" because it has not been tested against any possible regime shifts of a full annual cycle. There is no evidence presented that the model can "evolve" to handle the volatility of semi-volatile species in warmer months without catastrophic forgetting or significant error.

I strongly encourage the authors to extend their experiment to cover a longer period to genuinely establish the robustness of the incremental learning mechanism. Otherwise the authors need to rescale their claims. For example, the term "self-evolving" should be removed as the system's evolutionary capability remains unproven beyond Feb-Mar

2022. The authors must also explicitly discuss the theoretical risks of deploying this approach in operational setting outside the training season.

Authors' response:

We sincerely thank the reviewer for the crucial and insightful comment. We fully agree with the reviewer's concerns from both machine learning and physics perspective. Conducting a year-long experiment is crucial for verifying the robustness of the incremental learning mechanism of machine learning models, especially for tree-based models with poor extrapolation capabilities.

a. Re-scaling the claims on "self-evolving"

We fully agree that a two-month experimental period is insufficient to robustly demonstrate "self-evolution" against the full scenarios of atmospheric variability and unprecedented extremes. However, acquiring a year-long hourly observation dataset of five key $PM_{2.5}$ chemical components across a wide spatial range is currently very challenging. To our knowledge, none of the popular reanalysis datasets are generated directly from long-term hourly observations of chemical components. As summarized in **Table R1**, the CAQRA-aerosol dataset was generated indirectly by assimilating ground-level hourly observations of traditional air pollutants. The TAP dataset was generated by fusing daily, monthly, and annual observations of chemical components. The CHAP dataset was generated by fusing daily measurements of four water-soluble inorganic ions. The chemical component fields in both CAMSA and MERRA-2 were generated indirectly by assimilating observations of aerosol optical depth.

Consequently, within the current constraints, we deployed our maximum feasible effort to conduct a two-month hourly measurement campaign at 33 sites. This campaign was designed for a representative period (February-March 2022) and region (Beijing-Tianjin-Hebei region) known for frequent pollution episodes, which directly supports the primary goal of proposing and validating a novel framework online coupling incremental machine learning and ensemble data assimilation. **In response, we have re-scaled the claims on "self-evolving" and replace the term "self-evolving" with more precise descriptions such as "incrementally updatable" in the revised**

manuscript. In the future work, we will extend our measurement campaigns covering a longer period to establish the robustness of the incremental learning mechanism.

b. Discussion on extrapolation risks

In response, we have added a section ***3.4 Limitations*** to the revised manuscript to discuss the theoretical risks of deploying this approach in operational settings outside the training season. The original section ***3.4 Comparison with multiple reanalysis datasets*** has been changed to a section ***3.3.3. Comparison with multiple reanalysis datasets***.

3.4 Limitations, Line 621-637: "*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."

Table R1. The brief description of the observations used in the reanalysis datasets.

Dataset	Reanalysis species	Observed species	Temporal resolution of observation	Citation
CAQRA-aerosol	SO_4^{2-}, NH_4^+, NO_3^-, OC, BC	$PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, O_3	Hourly	Kong et al., 2025
TAP	SO_4^{2-}, NH_4^+, NO_3^-, OM, BC	SO_4^{2-}, NH_4^+, NO_3^-, OC, BC	Daily, monthly, and annual	Liu et al., 2022
CHAP	SO_4^{2-}, NH_4^+, NO_3^-, Cl^-	SO_4^{2-}, NH_4^+, NO_3^-, Cl^-	Daily	Wei et al., 2023
CAMSRA	NO_3^-, NH_4^+	Satellite-based AOD	12-hourly	Inness et al., 2019
MERRA-2	SO_4^{2-}, OM, BC	Satellite & ground-based AOD	Hourly	Randles et al., 2017

Reference

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.

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.

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.

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.

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.

Randles, C. A., da Silva, A. M., Buchard, V., Colarco, P. R., Darmenov, A., Govindaraju, R., Smirnov, A., Holben, B., Ferrare, R., Hair, J., Shinozuka, Y., and Flynn, C. J.: The MERRA-2 aerosol reanalysis,

1980 onward. Part I: System description and data assimilation evaluation, J. Clim., 30, 6823-6850, https://doi.org/10.1175/JCLI-D-16-0609.1, 2017.

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.

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.

The revised figures

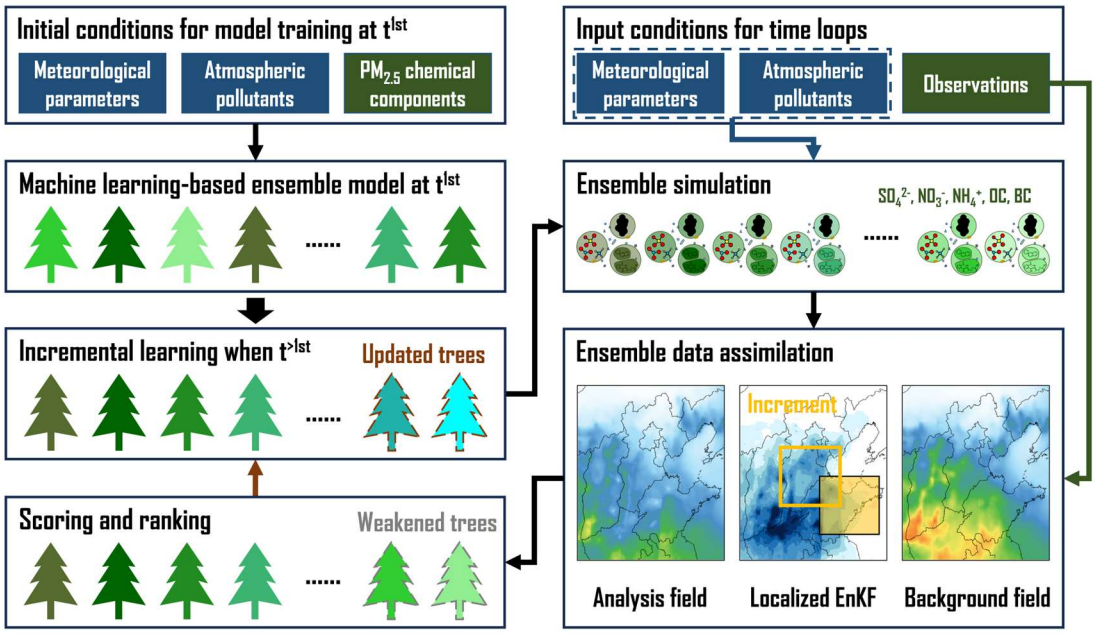

Figure 1. The framework of OIRF-LEnKF v1.0.

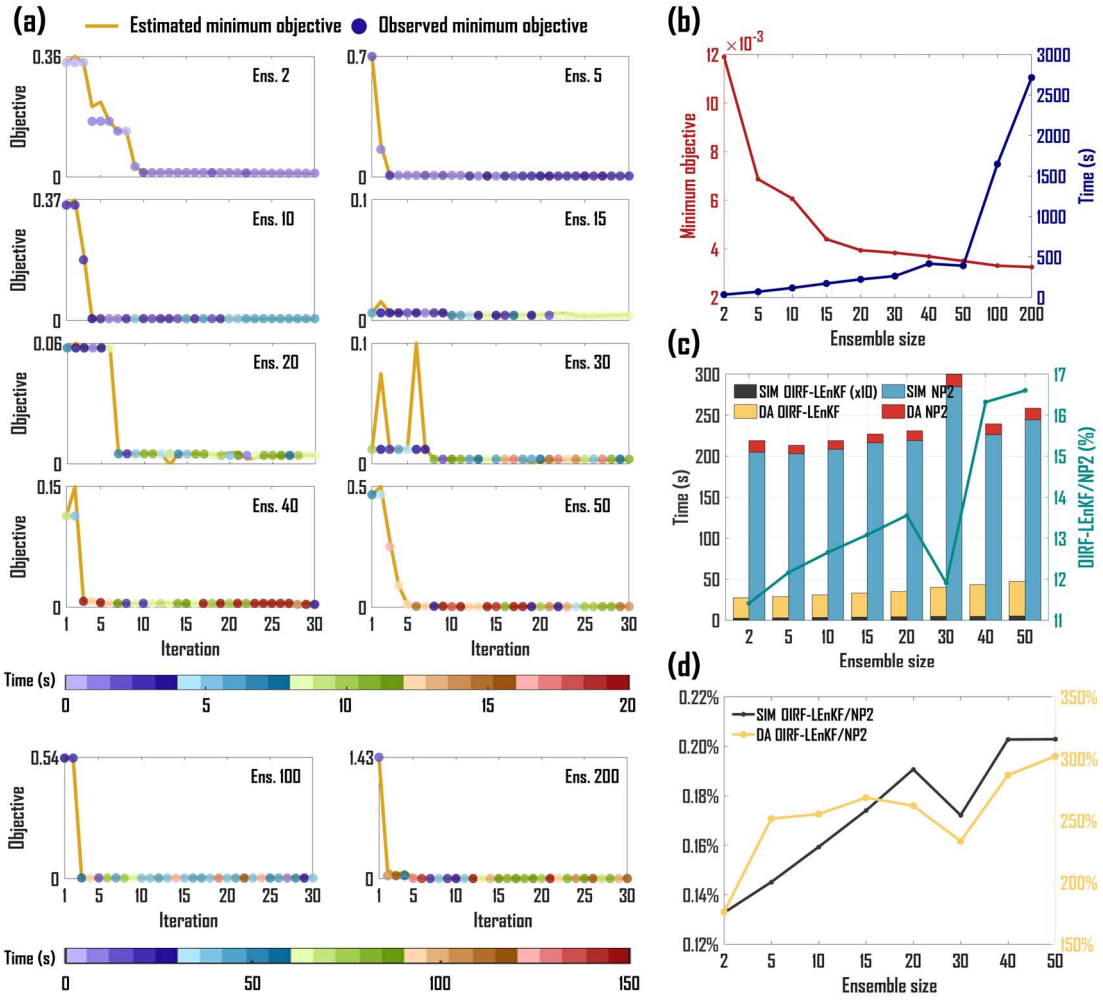

Figure 3. Computational efficiency of OIRF-LEnKF v1.0. (a) Variation in the minimum objective value throughout the Bayesian optimization process and time consumed by each iteration, determined by Eq. (5). (b) minimum value of total observed minimum objectives and total time consumed during Bayesian optimization 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 Figure 3c has been magnified by a factor of 10 for better clarity.

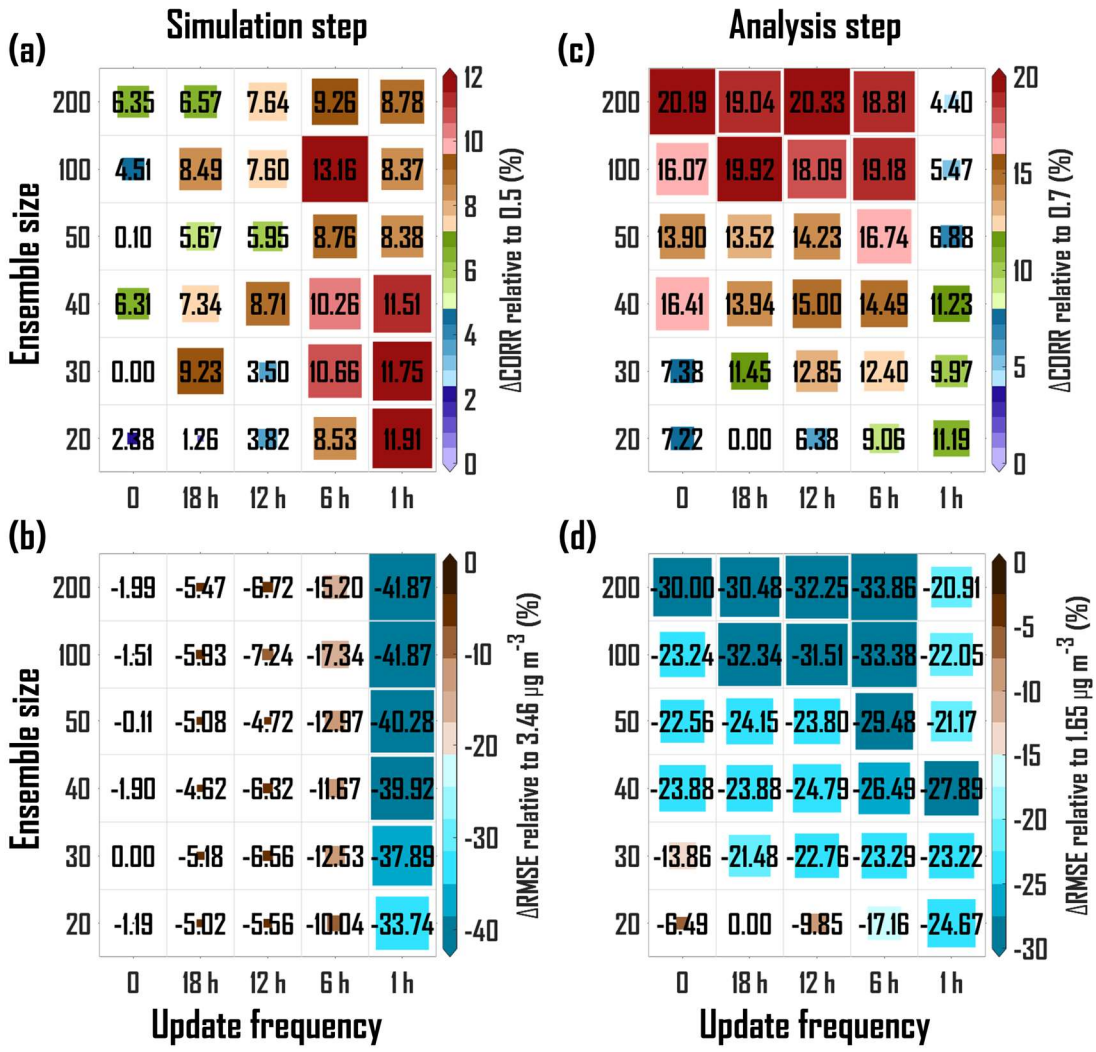

Figure 4. (a) Percentage change of Pearson correlation coefficient (CORR) relative to the minimum CORR (0.5) (ΔCORR, %) for sensitivity test with six ensemble sizes (20, 30, 40, 50, 100, 200) and five update frequencies (no update, 18-hour interval, 12-hour interval, 6-hour interval and 1-hour interval) 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.

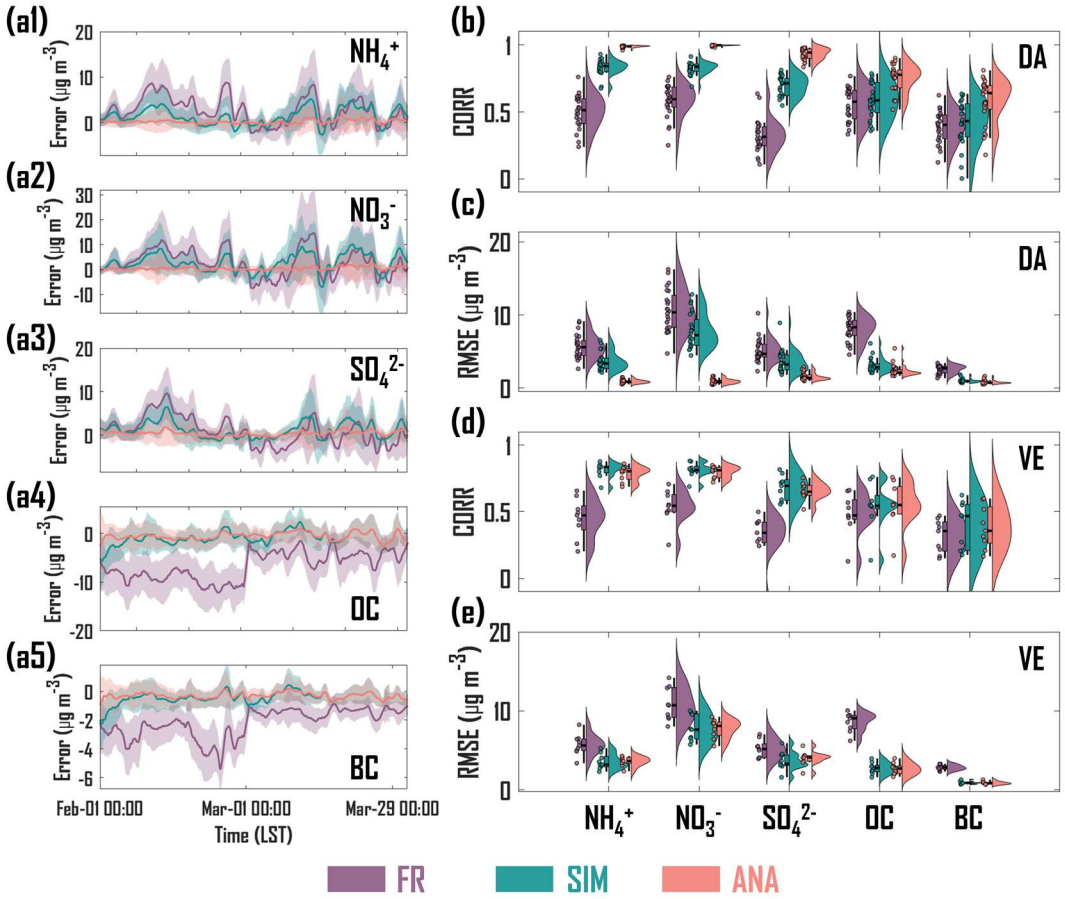

Figure 5. Smoothed variation in the error between observation and model output (including the free-run field (FR), the ML-simulated background field (SIM) and the analysis field (ANA)) for (a1) NH_4^+, (a2) NO_3^-, (a3) SO_4^{2-}, (a4) OC and (a5) BC at total sites during February and March of 2022. 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.

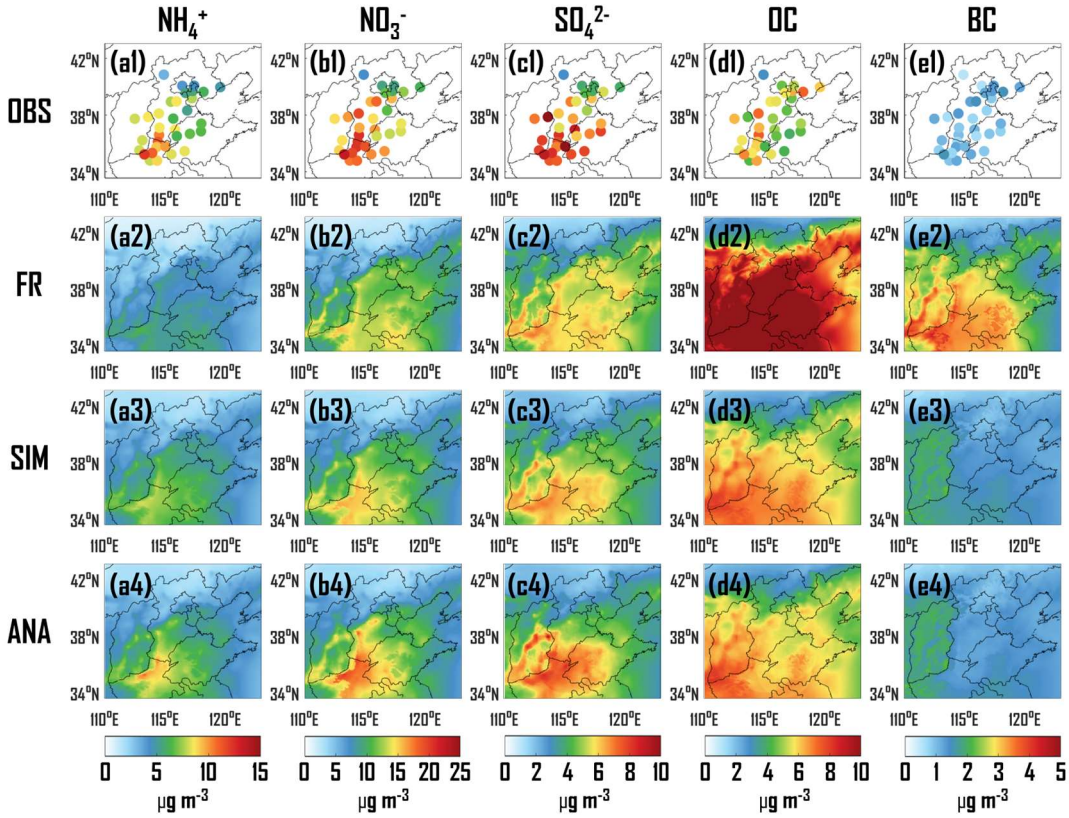

Figure 6. Spatial distribution of observation (OBS), free-run field (FRFR), ML-simulated background field (SIM) and analysis field (ANA) for NH_4^+ (a1-a4), NO_3^- (b1-b4), SO_4^{2-} (c1-c4), OC (d1-d4) and BC (e1-e4).

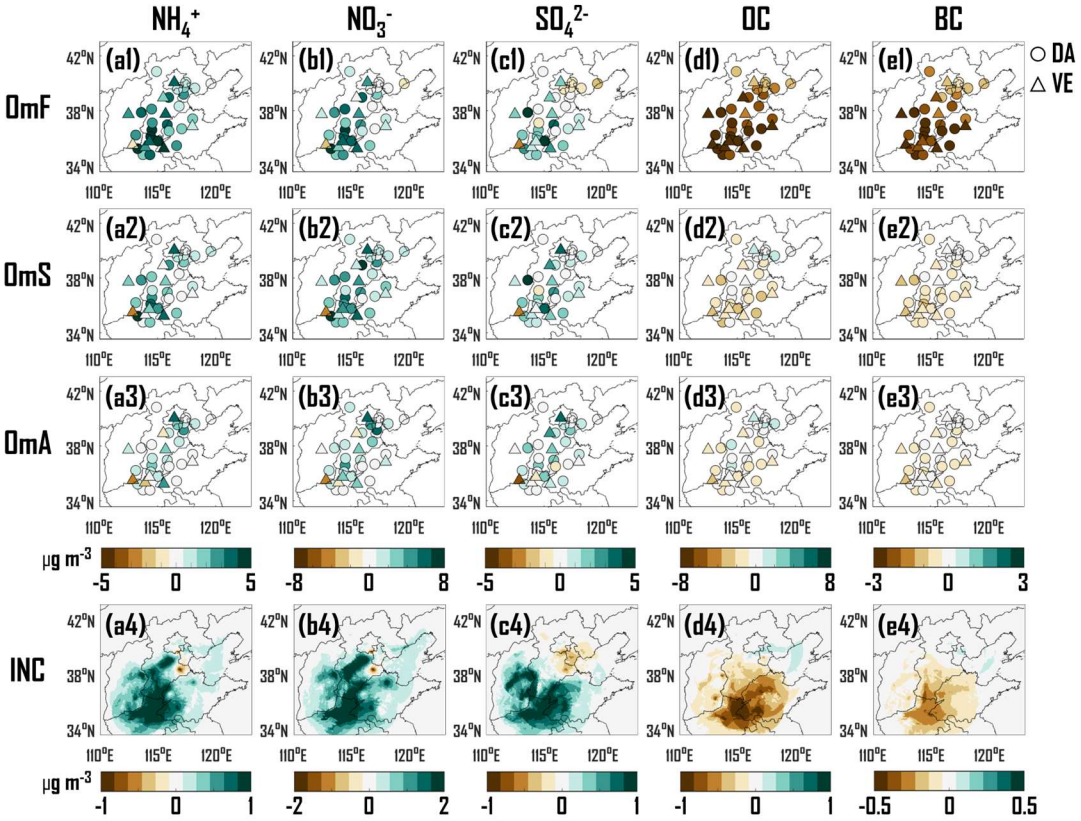

Figure 7. Spatial distribution of observation minus free-run field (OmF), observation minus ML-simulated background field (OmS), observation minus analysis field (OmA) and analysis field minus background field (INC) for NH_4^+ (a1-a4), NO_3^- (b1-b4), SO_4^{2-} (c1-c4), OC (d1-d4) and BC (e1-e4). The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