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

Geoscientific Model Development
Open Access
EGU

Supplement of

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 et al.

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

Contents of this file

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.

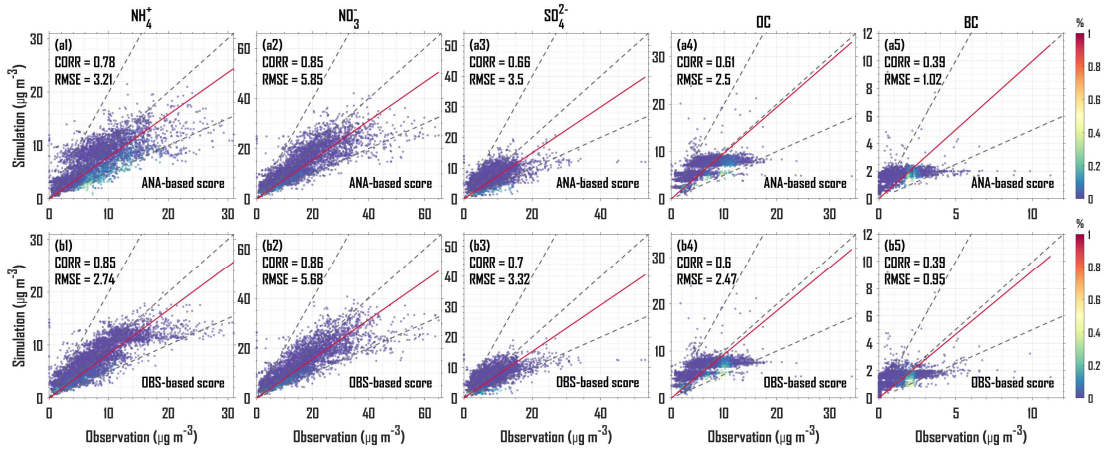

Figure S1: Scatterplots with probability density of simulated versus observed mass concentrations at independent VE sites correspond to the two scoring targets used in the incremental learning process, including analysis fields (ANA) (a1-a5) and independent observations (OBS) (b1-b5). The gray dotted lines represent the 2:1, 1:1, and 1:2 lines, and the red solid line represents the fitting regression line.