Referee Report: OIRF-LEnKF v1.0

Anonymous Reviewer

November 2025

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 PM2.5 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.

Major Comments

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 concurrent 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.

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.

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.