Response to Reviewer #2:

We sincerely thank the reviewer for the careful reading and constructive suggestions, which have helped improve the clarity and rigor of the manuscript.

Major revisions overview:

We have made several fundamental changes to the experimental framework in response to the reviewer's concerns about experimental scope and generalizability:

(1) Removal of the 9 km grid experiments. In the original submission, the 9 km grid served only as a supplementary test for computational efficiency analysis, while all core experiments—idealized tests, initial field scatter analysis, and forecast evaluation—were conducted exclusively at 27 km resolution. Given that the 9 km experiments were peripheral to the main findings, we removed them entirely in the revision. This allowed us to focus computational resources on substantially expanding the 27 km experiments, as detailed below.

(2) Extension from a single case to 20 consecutive days. The real-case assimilation experiment has been expanded from a single time step (00:00 UTC, October 27, 2023) to 20 consecutive daily cycles (October 10–29, 2023), each followed by a 24-hour forecast. This extension covers diverse meteorological and pollution conditions, enabling robust statistical evaluation of both assimilation accuracy and forecast skill.

(3) Addition of multi-core CPU scalability experiments. To address the reviewer's question about the performance bottleneck and parallelization efficiency, we conducted systematic scalability tests of both PY3DVar and FT3DVar using 1, 2, 4, 8, 16, 32, and 64 CPU cores across all 20 assimilation cases. This experiment provides a direct comparison of how the two systems scale with increasing parallel resources.

(4) Adoption of an independent validation strategy. The quality-controlled observations are now divided into assimilation and validation subsets: the gridded and thinned observations (averaging 546 stations, 28%) are used for assimilation, while the remaining observations (averaging 1,402 stations, 72%) are reserved exclusively for independent validation. All evaluation metrics are computed on the validation dataset, ensuring an unbiased assessment.

(5) Platform change: DCU to GPU, Intel to AMD. The DCU platform has been replaced by an NVIDIA RTX 3090 GPU (24 GB), which is a more widely accessible and familiar computing platform for the international readership. Similarly, the CPU platform has been updated from the originally listed Intel processor to a dual AMD EPYC 7542 configuration (64 physical cores @ 2.9 GHz, 256 GB memory), with the correct processor model now clearly documented.

(6) Substantial restructuring of the Introduction and Methods sections. The Introduction has been restructured to provide a more balanced and internationally representative literature review, with expanded coverage of recent advances. The Methods section has been reorganized, with a new dedicated section (Section 2.7) describing both assimilation systems and their implementation differences. All figures have been refined for clarity, and the conclusions have been revised to accurately reflect the scope of the study.

We address each specific comment in detail below:

> Comment:

The correlation, RMSE, and MAE statistics listed in the abstract are not quite useful. The paper shows computational efficiency improvement of the Py3DVAR over the Fortran-3DVAR, but there is little difference between the assimilation results of the two systems. So, the use of PyTorch does not improve the model initial fields. In addition, the assimilated observations are used in the evaluation. If possible, please use some independent data sets for evaluations. Otherwise, leave some observations from the assimilation and keep them for evaluation only..

****Response:****

We sincerely thank the reviewer for these valuable comments. We have revised the manuscript point by point as follows:

1. Regarding the statistical metrics in the abstract:

We fully agree with your opinion. The correlation, RMSE, and MAE values have been completely removed from the abstract in the revised version.

2. Regarding the assimilation performance of PY3DVar and FT3DVar:

We appreciate the reviewer's observation that the assimilation results of the two systems show negligible differences. We would like to clarify that the primary objective of this study is not to achieve better assimilation accuracy than the traditional Fortran system. The existing FT3DVar has already demonstrated reliable and satisfactory assimilation performance in practical applications. PY3DVar, built upon the PyTorch framework, is designed to reproduce equivalent

assimilation accuracy while achieving substantial computational acceleration. Therefore, the close agreement between the two systems is exactly the result we pursue—it verifies that the newly developed PY3DVar can fully match the performance of the mature Fortran counterpart, ensuring its reliability and readiness for operational adoption. We have clarified this core motivation in the revised manuscript.

3. Regarding dataset separation for evaluation:

We fully acknowledge the reviewer's suggestion to use independent datasets for evaluation. However, due to the limited availability of independent observation networks (e.g., satellite retrievals or radiosonde profiles) covering the study domain and time period, we instead adopted a strict data-thinning and separation strategy to prevent evaluation bias. As described in Section 2.2, the quality-controlled observations are divided into two mutually exclusive subsets: the thinned observations (averaging 546 stations, 28%) are used in the assimilation process, while the remaining observations (averaging 1,402 stations, 72%) are reserved exclusively for independent validation (Line228-232). This separation ensures that the evaluation metrics are computed solely from observations that were not involved in the assimilation, which is a widely accepted practice in data assimilation studies when independent external datasets are unavailable. We have added a detailed description of this separation strategy to the revised manuscript.

> Comment:

It is highly suggested that the authors extend their testing beyond the single day they chose here.

**** Response:****

Thank you for this valuable suggestion. We fully agree that multi-day experiments can better validate the stability and general applicability of the assimilation system.

In the revised manuscript, we have extended the experimental period significantly. All real-case assimilation and forecast experiments are performed continuously from 00:00 UTC on 10 October 2023 to 00:00 UTC on 29 October 2023, covering 20 consecutive daily cycles, rather than a single day. Over this long time series, we comprehensively evaluated the assimilation quality(Section 4.2), numerical stability, parallel scalability(Section 4.4) and computational efficiency (Section 4.3, 4.4) of both PY3DVar and FT3DVar. The multi-cycle tests under continuously changing environmental conditions fully prove that the conclusions drawn in this study are robust and not limited to a single momentary case.

> Comment:

In Introduction, the papers selected of chemical data assimilation applications are not very

representative of the advancement of the field. The selections seem to be biased toward the works from their own region.

****Response:****

We sincerely thank the reviewer for this critical and constructive comment. We fully agree that the original Introduction was unbalanced and relied too heavily on regional studies. To address this, we have substantially revised the literature review to include a broader and more internationally representative set of references, covering foundational work, recent methodological advances, and applications from diverse research groups across Europe, North America, and Asia. The key additions to the revised Introduction include:

(1) Foundational chemical data assimilation studies:

• Europe: Elbern et al. (1997) [EURAD system]; Denby et al. (2008); Frydendall et al. (2009); Tombette et al. (2009).

• North America: Austin et al. (1992) [U.S. EPA framework]; Pagowski et al. (2010) [WRF-Chem assimilation].

• East Asia: Li et al. (2013) [China]; Jiang et al. (2013) [China].

(2) Recent advances in chemical data assimilation (2023–2025):

• Seo et al. (2023) – Ensemble-based approaches in East Asia.

• Chen et al. (2024) – Deep learning-enhanced assimilation frameworks.

• Zhou et al. (2025) – Aerosol data assimilation using satellite retrievals.

• Ma et al. (2025) – Advances in variational assimilation over China.

(3) Novel methodological developments from diverse international groups:

• Cheng et al. (2025) – PyTorch-based DA package development.

• Key et al. (2024) – Scalable DA with message passing.

• Chattopadhyay et al. (2023) – Machine learning surrogate models for DA.

• Beauchamp et al. (2025) – Nonlinear Bayesian filtering methods.

• Xu et al. (2025) – Hybrid variational-ensemble techniques.

(4) Classic theoretical foundations:

• Lorenc (1986) – Statistical theory of 3DVar.

• Parrish and Derber (1992) – NMC method for background error covariances.

We believe the revised Introduction now provides a representative and up-to-date overview of the field's advancement. The corresponding changes have been made in the revised manuscript. We sincerely appreciate the reviewer's guidance in helping us broaden the scope and quality of our literature coverage.

> Comment:

Many figures are pretty difficult to read.

****Response:****

We sincerely appreciate the reviewer's constructive feedback regarding the readability of the figures. To address this concern, we have carefully revised and expanded the figure captions and the corresponding text in the manuscript. Below, we provide a detailed, step-by-step explanation of how each complex figure was generated, the meaning of the axes and graphical elements, and the scientific purpose they serve.

****Figure 2. Vertical profiles of background error standard deviation.****

- **Data source and calculation:** The background error standard deviation matrix D is derived using the NMC method (Parrish et al., 1992). We performed 24-hour and 48-hour WRF-Chem forecasts daily at 00:00 UTC over 30 consecutive days. The forecast difference fields $\varepsilon = X\_48h - X\_24h$ are computed. The standard deviation for each grid point and control variable is then calculated from these 30 daily difference samples (Equation 6 in Section 2.6) and averaged horizontally to produce the 1D vertical profiles shown in Figure 2.
- **Axes:** The left y-axis represents the model's vertical level (from level 1 near the surface to level 34 at the top of the atmosphere). The right y-axis shows the corresponding physical height. The x-axis represents the background error standard deviation.
- **Graphical elements:** Each colored line represents a specific chemical or aerosol species.
- **Physical meaning and purpose:** This figure quantifies the uncertainty of the model background field at different altitudes. A larger standard deviation (e.g., for DUST species) indicates higher uncertainty in the background simulation, which means the assimilation system will apply larger corrections (analysis increments) to these variables. Conversely, a smaller standard deviation implies higher confidence in the background, resulting in minor corrections. The figure demonstrates that uncertainty generally decreases with height for most species, while O_3 shows an increasing trend due to its longer lifetime in the upper troposphere.

****Figure 3. Distribution of the vertical autocorrelation matrix of background error.****

- **Data source and calculation:** The vertical correlation matrix Cz is computed from the same 30-day forecast difference fields. For each species, we calculate the correlation coefficients between all pairs of vertical layers (Equation 7 in Section 2.6). The resulting values are averaged over all horizontal grid points to produce the 34*34 matrix shown in Figure 3.
- **Axes:** Both axes represent the vertical model levels (1 to 34). The color intensity (or

shading in the matrix plots) represents the value of the vertical correlation coefficient, ranging from 0 (no correlation) to 1 (perfect correlation).

- **Physical meaning and purpose:** This figure reveals how errors in the background field are vertically correlated. Because strong near-surface correlation is observed (values close to 1 near the diagonal at lower levels), the assimilation system effectively propagates surface observational information upward to correct the lower atmosphere. The decreasing correlation with height indicates that observational data has a stronger constraint on the lower levels than on the upper levels.

Figure 4. Decay of horizontal background error correlation coefficient with distance.

- **Data source and calculation:** The horizontal correlation scale parameter L is estimated using the same forecast difference fields. We compute the correlation coefficient d between all pairs of grid points and group them according to their horizontal separation distance d. The curves in Figure 4 represent the empirical average of these correlation coefficients at each distance bin, rather than a fitted mathematical function (as explicitly stated in the revised text).

- **Axes:** The x-axis represents the physical horizontal distance in kilometers (km), derived from the model's grid spacing of approximately 27 km. The figure is plotted only up to 500 km because correlations beyond this distance are negligible (well below the $e^{-1/2}$ threshold). The y-axis represents the horizontal correlation coefficient, ranging from 0 to 1.

- **Graphical elements:** Each curve represents the decay rate for a specific variable. The red dashed horizontal line marks the $e^{-1/2}$ (approximately 0.6065) threshold.

- **Physical meaning and purpose:** This figure determines the horizontal influence radius of observational data in the assimilation system. The scale parameter L is defined as the maximum distance where the correlation coefficient remains above the $e^{-1/2}$ threshold. Because different species exhibit different decay rates, the calculated L values range from 1 to 4 grid points. This heterogeneity governs how widely observations for each pollutant are spread horizontally during the assimilation process.

Figure 12. Statistical characteristics of iteration performance for different assimilation systems.

- **Data source and calculation:** The data in Figure 12 comes from 20 consecutive days of real-case atmospheric assimilation experiments. (a) The number of iterations required for convergence is recorded for each daily cycle for both FT3DVar and PY3DVar. (b) Total iteration time per cycle is the elapsed wall-clock time from the start to the end of the minimization process.

(c) Mean per-iteration time is calculated by dividing the total iteration time by the number of iterations for each cycle. (d) The cost function value is recorded after every minimization step; the curves are averaged over the 20 days.

- **Graphical elements:**

  - **Fig. 12a:** Bar plot showing the number of iterations per day. Blue bars = FT3DVar; Red bars = PY3DVar.

  - **Fig. 12b:** Line plot showing the total time per day. Blue line = FT3DVar; Red line = PY3DVar.

  - **Fig. 12c:** Line plot showing the mean time per iteration. Blue line = FT3DVar; Red line = PY3DVar.

  - **Fig. 12d:** Line plot showing the mean cost function descent curve. The inset panels zoom in on specific iteration intervals (Iter 4–7 for PY3DVar and Iter 16–23 for FT3DVar) to show the detailed convergence behavior at the late stages.

- **Physical meaning and purpose:**

  - **Panels (a), (b), and (c)** work together to explain *why* PY3DVar is faster overall: Although PY3DVar takes longer per iteration than FT3DVar (Fig. 12c), it requires substantially fewer iterations to converge (Fig. 12a). This reduction in iteration count overcomes the per-iteration time penalty, resulting in a significantly shorter total iteration time (Fig. 12b).

  - **Panel (d)** illustrates the L-BFGS convergence behavior. The steeper drop in the PY3DVar curve during the first 4 iterations reflects the Strong Wolfe line search strategy (used in PyTorch), which yields a larger cost function reduction per step compared to the Moré-Thuente scheme used in FT3DVar. The closely aligned final cost values confirm that both systems achieve the same assimilation accuracy.

****Figure 13. Parallel iteration time comparison across different CPU core counts.****

- **Data source and calculation:** The data is generated by running the 20-day assimilation cycles on identical CPU hardware, but using different core counts (1, 2, 4, 8, 16, 32, and 64). The total iteration time and per-iteration time are recorded and averaged over the 20 days for each core configuration. The speedup in Figure 13c is calculated as the ratio of PY3DVar's total iteration time to FT3DVar's total iteration time.

- **Graphical elements:**

  - **Fig. 13a:** Bar plot comparing mean per-iteration time.

  - **Fig. 13b:** Bar plot comparing mean total iteration time.

  - **Fig. 13c:** Line plot showing the speedup curve. The red dashed horizontal line indicates a

speedup of 1 (the "parity line" where both systems take the same time).

- **Physical meaning and purpose:** This figure analyzes the trade-off between per-iteration speed and parallel scalability. As shown in (a), FT3DVar is inherently faster per iteration on all CPU core counts. However, PY3DVar exhibits much better parallel scaling (the time drops rapidly as cores increase) and requires fewer iterations. As a result, as shown in (b) and confirmed by the speedup curve in (c), PY3DVar surpasses FT3DVar in total time when 32 cores or more are used. The inflection point where the speedup curve crosses 1.0 (between 16 and 32 cores) clearly marks the transition where PY3DVar becomes more computationally efficient overall.

We believe these detailed explanations, now fully integrated into the figure captions and the main text, will substantially improve the readability and accessibility of our figures. We have also refined the visual design of the plots (e.g., enlarged font sizes, clearer legends) to further aid comprehension. For the remaining figures in the manuscript, such as the scatter plots and spatial distribution maps, we believe their axes, legends, and physical meanings are sufficiently self-explanatory based on the standard conventions of atmospheric science and do not require further detailed elaboration here. We thank the reviewer for helping us improve the clarity of our presentation.

> Comment:

The descriptions of the idealized tests can be shortened.

****Response:****

We have substantially shortened the descriptions of the idealized experiment as suggested. The revised text retains core results and key conclusions while removing redundant content, making the presentation more concise and focused (Line 464-473).

Specific comments:

> Comment:

Lines 17-21, “In real-case experiments, Py3DVAR substantially improves the model initial field quality: at 27 km resolution, correlation coefficients …”: This could be misleading since Py3DVAR and Fortran-3DVAR perform similarly.

****Response:****

We removed the standalone correlation, RMSE and MAE values from the original abstract to

avoid overemphasizing analysis improvements. The updated abstract clearly states that PY3DVar and FT3DVar yield nearly identical chemical analysis accuracy, which clarifies that the core advancement of our new system lies in computational acceleration rather than superior assimilation skill.

> Comment:

Lines 68-69, "… employing relatively simple methods. For example, Denby et al. (2008) applied both Statistical Interpolation (SI) and Ensemble Kalman Filtering (EnKF) data assimilation methods …": EnKF is not considered a simple method.

****Response:****

Thank you for this precise correction. The expression "relatively simple methods" was inappropriate, as EnKF is a sophisticated ensemble-based assimilation approach rather than a simple technique. We have revised the corresponding sentence in the manuscript to remove this inaccurate description (Line 68).

> Comment:

Lines 217-218, "cost function tensor operations": What does it mean here? Cost function is not a tensor.

****Response:****

We sincerely thank the reviewer for pointing out this inaccurate terminology. We fully agree that the cost function itself is a scalar, not a tensor, and that the phrase "cost function tensor operations" was mathematically inappropriate.

In the revised manuscript, this ambiguous wording has been removed. As described in Section 2.5, we have clarified the computational implementation as follows: all core computations—including the background error covariance transformations, the observation operator, and the cost function evaluation—are formulated as tensorized matrix operations within the PyTorch framework (Line 322-324). The cost function minimization is performed by processing these underlying matrix and vector calculations using tensor operations, without implying that the scalar cost function itself is a tensor.

We appreciate the reviewer's careful reading, which helped us eliminate this mathematical ambiguity.

> Comment:

Line 260, "vertical configuration of 34 layers": The vertical extent and grid spacings need to be

explained.

Response:

We thank the reviewer for this valuable comment. In the revised manuscript, we have added a comprehensive explanation of the vertical structure (Lines 185–187). The model employs 34 terrain-following vertical layers, with higher resolution near the surface that gradually decreases with increasing altitude. The vertical domain extends from the surface up to approximately 20 km. Additionally, the domain-averaged height of each vertical level is provided in Fig. 2 to clearly illustrate the vertical distribution characteristics.

> Comment:

Equation (8): The horizontal correlation scale parameter L does not have to be an integer. However, Figure 4 shows L are all integers. Please explain whether L are chosen to be integers only. If so, why?

Response:

Thank you for the valuable comment. The values of the horizontal correlation scale parameter L shown in Fig. 4 are indeed integers, and we clarify the rationale below.

In this study, L is defined directly from the empirical correlation decay curve as the maximum grid distance d at which the correlation coefficient $\rho(d)$ remains $\rho(d) \geq e^{-1/2}$ (see Equation 8). Since d is inherently a discrete quantity (measured in integer grid spacings), the resulting L is naturally an integer. This definition is deliberately non-parametric: it extracts L directly from the sample statistics without invoking any intermediate fitting procedure.

An alternative approach, adopted in some previous studies (e.g., Wang et al., 2022, GMD: https://doi.org/10.5194/gmd-15-1821-2022), is to fit the empirical correlation curve via least-squares regression, thereby obtaining a continuous (non-integer) estimate of L. While this yields a more "precise" numerical value, it introduces additional uncertainty through the choice of the fitting algorithm.

Our integer-based definition offers several practical advantages in the context of operational 3DVar:

1. Simplicity and robustness: It avoids the sensitivity of non-linear least-squares fitting to initial guesses and local minima, which can be problematic with noisy empirical correlation profiles.

2. Physical interpretability: L directly corresponds to the number of grid points within which the background error correlation exceeds the $e^{-1/2}$ threshold, providing an intuitive measure of the influence radius of observations.

3. Computational consistency: Since the horizontal correlation matrices C_x and C_γ are constructed

on the discrete model grid, an integer L is fully consistent with the discrete nature of the numerical implementation.

> Comment:

Line 517, “observation intensities”: It is not clear what the authors mean by “observation intensities”. Please clarify.

****Response:****

We apologize for the ambiguous expression of "observation intensities" in the original text. We have now replaced this phrase with different concentration magnitudes throughout the revised manuscript (Line 435).

This term specifically refers to the two distinct levels of prescribed pollutant concentrations adopted in the idealized experiment, namely the low-concentration and high-concentration observation settings. The revision clearly indicates that we examine the performance of the assimilation system under two pollutant concentration levels. The relevant sentence and wording have been updated accordingly.

> Comment:

Figures 7 and 8: It is better to use the actual altitudes rather than “Vertical Level” numbers in these figures.

****Response:****

Thanks for the valuable suggestion. We have fully revised the axis labels of Figs. 7 and 8. The vertical axis originally marked with vertical level indices is now replaced with actual altitudes. Meanwhile, the horizontal axis, which previously used grid point indices, has been updated to actual horizontal distances calculated based on the 27 km grid spacing. These revisions make the spatial features of vertical cross-sections more intuitive and quantitatively explicit.

> Comment:

Line 740, “Hessian matrix approximation”: What is the chosen “m” value for the L-BFGS routine?

****Response:****

We thank the reviewer for this comment. In the revised manuscript, we have thoroughly restructured the relevant analysis section, and the original description of the Hessian matrix approximation has been entirely removed. For completeness, we clarify the L-BFGS parameter settings used in both PY3DVar and FT3DVar: the memory size (i.e., the number of stored history

pairs, typically denoted as m) is uniformly set to 6, and the maximum number of inner iterations per optimization step is set to 4 (Line 299-303). These settings are identical for both implementations.

Editorial comments:

> Comment:

Line 16, “Idealized results” should be “The results of the idealized tests”.

****Response:****

We have revised the phrase at Line 16 from "Idealized results" to The results of the idealized tests as suggested, to conform to standard academic writing conventions (Line 18).

> Comment:

Line 129, “Wang et al.” -> “Wang et al. (2022)”

****Response:****

We thank the reviewer for this comment. The original citation at Line 129 has been removed in the revised manuscript. Additionally, we have systematically checked and corrected all remaining "Wang et al." citations in the manuscript by adding the publication year (e.g., Wang et al., 2022) to ensure consistency with the journal’s reference format.

> Comment:

Line 509, consider using “with” instead of “under” .

****Response:****

We have followed this suggestion and replaced the preposition "under" with "with" in the corresponding sentence (Line 428).

> Comment:

Line 593, “Fig. 7and 8” -> “Figs. 7and 8”

****Response:****

We have corrected the abbreviation from "Fig. 7and 8" to "Figs. 7 and 8" (Line 465).

> Comment:

Line 703, “narrows” -> slows

****Response:****

Thank you for the comment. We have fully restructured the corresponding section, and the sentence containing the word "narrows" has been completely removed in the revised manuscript.

> Comment:

Lines 956-958: This paper is out of place and should be moved down.

****Response:****

We have adjusted the position of the referenced content as suggested and moved these lines to a more appropriate section in the manuscript.