This manuscript uses f2py to couple machine-learning-based latent data assimilation (LDA) with traditional numerical models. The results suggest that the proposed hybrid system can run stably and achieves performance comparable to conventional methods. The topic is timely and potentially valuable, as LDA has shown promising advantages over the model-space DA methods, yet it has rarely been deployed within traditional numerical models. However, the manuscript currently suffers from substantial issues in terms of topic framing, experimental design, algorithmic clarity, and writing quality. Overall, I believe the study has potential, but major revisions are required before it can be considered for publication.

Comments:

1. The current scope and framing of the manuscript are not well aligned with its actual contribution.

(1) The title emphasizes *AI and Physics Modeling and Data Assimilation*, but the manuscript is in practice mainly concerned with coupling ML-based LDA with traditional numerical models.

(2) The introduction highlights the promise of AI for physical modeling, the main text does not present a concrete example in which AI is used to improve the physical model itself. In addition, the introduction does not clearly explain why ML-based DA is needed in this setting.

(3) Section 2 mainly describes Python–Fortran interfacing and f2py-based wrapping in a fairly general way, rather than explaining how AI-based physical modeling and data assimilation are implemented within the numerical model.

I suggest that the manuscript be reframed so that its main focus is clearly defined as the integration of LDA into traditional numerical modeling systems. In particular, Section 2 should focus more on how LDA is implemented and integrated with the traditional model. LDA is an algorithmic framework, and its current description is too vague.

2. The description and implementation of the LDA algorithm require substantial clarification and revision.

(1) The description of the algorithm is too limited to support reproducibility. For example, it is unclear how the latent-space background error covariance is specified or estimated.

(2) The manuscript should explicitly state that the method used here is Latent-3DVar (L3DVar), since LDA is a broader framework that includes multiple algorithmic variants.

(3) The LDA equations should be presented in the main text (Section 2) rather than embedded in a figure.

(4) The notation for the background error covariance is conceptually inconsistent. In Fig. 5, B is defined as the model-space background error covariance matrix, but the same notation is also used for the latent-space background error covariance in the LDA formulation. To avoid confusion and remain consistent with the

current LDA work, the latent-space covariance should be denoted by B_z.

(5) A key advantage of applying LDA to high-dimensional atmospheric DA is that the B_z can become approximately diagonal in latent space. By checking the code, I found that the manuscript sets this covariance to the identity matrix empirically, which may be problematic. The authors should revise the treatment of B_z in their experiments and clearly explain how it is estimated in the manuscript. I recommend estimating B_z with NMC method or from ensembles.

(6) The citation of LDA is incomplete. To my knowledge, [1] is the first study showing that B_z is nearly diagonal in a high-dimensional atmospheric setting, and [2] is the first extending LDA to multivariate atmospheric DA while also showing the same property. In my view, both references should be cited. Moreover, the citation to Fan around Line 245 should refer to Part 2 ([4]) rather than Part 1 ([3]).

(7) In Section 4, the autoencoder is trained using forecast data rather than the reanalysis data more commonly used in LDA studies. This choice should be explained more clearly. I suggest referring to [2] and [3], which both use forecast data to train the autoencoder.

(8) The authors should not refer to their model as a beta-VAE. In the standard usage, a beta-VAE is a generative model with an increased weight on the KL term (i.e., KL weight > 1). In the present case, it would be more appropriate to refer to the model simply as a VAE, or more accurately as an AE with slight KL regularization. I consider the latter description to be the most precise.

(9) As discussed in [2], the validity of L3DVar relies on conditions such as the approximately affine behavior of the decoder. The authors should clarify this point in the manuscript.

[1] B. Melinc, Ž. Zaplotnik, 3D‐Var data assimilation using a variational autoencoder. Quarterly Journal of the Royal Meteorological Society 150, 2273–2295 (2024).

[2] H. Fan, L. Bai, B. Fei, Y. Xiao, K. Chen, Y. Liu, Y. Qu, F. Ling, P. Gentine, Physically consistent global atmospheric data assimilation with machine learning in latent space. Sci. Adv. 12 (2026).

[3] H. Fan, Y. Liu, Z. Huo, Y. Liu, Y. Shi, Y. Li, A Novel Latent Space Data Assimilation Framework with Autoencoder-Observation to Latent Space (AE-O2L) Network. Part I: The Observation-Only Analysis Method. Monthly Weather Review 153, 1335–1348 (2025).

[4] H. Fan, Y. Liu, Y. Liu, Z. Huo, B. Chen, Y. Qin, A Novel Latent Space Data Assimilation Framework with Autoencoder-Observation to Latent Space (AE-O2L) Network. Part II: Observation and Background Assimilation with Interpretability. Monthly Weather Review 153, 1349–1363 (2025).

3. The experimental design is too vague, and in my view the resulting conclusions may not be sufficiently

reliable, as the LDA method does not appear to have been used appropriately.

(1) In both cases, LDA is compared with EnKF, which is ensemble-based. However, I do not see a corresponding ensemble-based design for estimating the background covariance in the LDA implementation. This does not seem to provide a fully fair comparison.

(2) As noted above, the latent-space background covariance B_z does not appear to have been properly estimated, which may lead to poor LDA performance.

(3) In the WRF experiment, the compression ratio appears to be too small, which may mean that the background error covariance in latent space is not close to diagonal. In that case, using a diagonal B_z may introduce substantial assimilation error.

(4) The manuscript does not clearly explain how the observation error covariance R is defined.

(5) The paper should include a comparison of the computational efficiency of the conventional DA method and LDA.

(6) The manuscript does not provide sufficient statistical evaluation of the autoencoder reconstruction error, but only reports a few individual cases.

(7) The naming of the LDA algorithm should be kept consistent across the different experiments, or at least be made as consistent as possible. The current names for LDA methods are confusing.

4. The quality and presentation of the figures require substantial improvement in the revised manuscript.

(1) Font styles are inconsistent across figures, which gives the figure set an unpolished appearance. A unified font style and size scheme should be used throughout.

(2) Figures 7 and 8 appear to present case studies or example events. If so, the corresponding dates or timestamps should be explicitly labeled.

(3) Several figures are devoted to presenting AE reconstruction errors (e.g., Figures 7, 8, 12, and 13). Some should be moved to the Appendix.

(4) Figures 6 and 11 are difficult to read, with text that is too small and a layout that appears overly simplistic. The AE architecture is best presented graphically rather than described only in words. The authors can refer to the presentation style used in existing LDA studies.

(5) In some figures, the color bars are placed too tightly relative to the main panels, which affects readability. Figure 7 is one example.

(6) The color bars in Figure 9b and 9c appear to be incorrect and should be checked carefully.

(7) In Figure 10, the borders around the subpanels should be removed, and the subpanel labels should not be placed directly on the images with a white background box. The authors can refer to recently published

papers in this journal for more standard presentation styles.

(8) In my view, Figure 2 could be removed and its content briefly summarized in the text.

(9) Choose a color bar for showing difference fields (like “bwr”) in Figure 14.

(10) Overall, the visual quality and aesthetics of the figures should be improved.

5. The writing of the manuscript requires substantial improvement.

(1) Some expressions are uncommon in academic English. For example, at line 319, “From there, we can see that” could be replaced by “It shows that” or, preferably, by stating the conclusion directly. Similarly, at line 349, “From Figs. 9b-c, we can see that” could be revised to “Figs. 9b-c show that”.

(2) Some section titles are unclear or not written in standard academic English. For example, “3.3.3 The VAE’s data-training and latent space minimization” is awkward, and the distinction between “3.4 Preliminary results of SCDA” and “3.4.2 The SCDA results” is confusing.

(3) The wording needs to be more precise. For example, at line 248, “ensures” should likely be replaced by “encourages”, since the KL loss does not ensure the desired latent-space property.

(4) Some citations are missing or should be checked more carefully. For example, the first appearance of VAE around line 241 should be supported by a citation.

(5) There are too many abbreviations, and some of them are not sufficiently clear. I suggest reducing or consolidating some of the abbreviations.

(6) In the abstract, some abbreviations are used before being defined. For example, CDA is used before being explained, and f2py should also be briefly introduced.

The manuscript contains many issues in English expression. I have only listed a few representative examples here, and the authors should carefully check the language throughout the manuscript.