

RESPONSE TO THE REVIEWER'S COMMENTS

In the following, we provide detailed responses to the comments of the reviewers. For each comment, we first reproduce the reviewer's original comment, followed by our response and a description of the corresponding changes made to the manuscript. The original comments from the reviewers are reproduced in normal font, while our responses are shown in blue. References to sections, figures, and tables refer to the revised manuscript unless otherwise indicated.

General comments

- A) We thank both reviewers for their careful evaluation of the manuscript and for their constructive and helpful comments and suggestions. Their feedback has helped us improve the manuscript considerably. In response to the reviews, we have substantially revised the manuscript, expanding the scope of the work by adding new applications of the methodology and a new experiment based on a real volcanic eruption.

The original manuscript demonstrated the feasibility of expanding volcanic-cloud ensembles using generative AI under a controlled numerical experiment involving a hypothetical eruption. The revised manuscript extends this work in two important directions: (i) the inclusion of latent-space data assimilation and (ii) the application of the methodology to the 2018 Etna eruption.

The manuscript has also been substantially revised and restructured. The revised organization is as follows:

1. [Introduction](#)
2. [Variational Autoencoder](#)
3. [Latent-space data assimilation](#)
4. [Numerical experiments and evaluation](#)
5. [Results and discussion](#)
6. [Conclusions](#)

The original numerical experiments based on the hypothetical Garachico eruption have been retained and are now presented as the synthetic experiment. However, some changes to the original experiments have been introduced in the revised manuscript, as detailed below.

Main modifications to the original manuscript:
(here, references correspond to the original manuscript)

- **Title:** The manuscript title has been updated to “Integrating generative AI with physics-based simulations for probabilistic modelling of volcanic clouds” to better reflect the broader scope of the revised work.
- **Table 2 (Ensemble configuration)** has been removed from the main text and moved to the Supplementary Material. Detailed model configuration information is now provided in the Supplementary Material to reduce the amount of technical detail in the main text and avoid duplicating the table for the new experiment.
- **Table 3 (Datasets)** has been removed from the main text following Reviewer #1’s suggestion. The information on ensemble sizes has been incorporated into Table 1.
- **Figure 5 (Ensemble mean computed from the test dataset)** has been moved to the Supplementary Material, as it does not provide significant additional value to the main text.
- **Figure 7 (Exceedance probabilities)** has been updated in response to Reviewer #1’s suggestion, with revised colors and line styles to facilitate the identification and referencing of the different contour levels.
- **Figure 10 (Isotropic power spectrum)** has been updated by testing different decoder architectures to improve the discussion in the new Section 4.4 (Isotropic power spectrum), as requested by Reviewer #2.
- In addition, we provide **Supplementary Material** containing additional figures, links to the repository hosting the full source code used in this work, and the input configuration files required to reproduce the numerical experiments.

In summary, two tables have been removed from the main text or moved to the Supplementary Material, one figure has been moved to the Supplementary Material, and two new figures have been added for the Etna experiment (Figs. 10 and 11).

Finally, we have revised the manuscript throughout to make the potential applications of the methodology in real-world scenarios more evident. We have also moderated statements concerning operational applications and clarified that the present work should be regarded as a contribution to an emerging research area, providing a basis for further investigation rather than a ready-to-use operational forecasting methodology.

Given these substantial extensions, we believe that the revised manuscript goes beyond the scope of a purely methodological technical note. We therefore respectfully encourage the editor to reconsider the current Technical note designation and to consider the revised manuscript as a standard research article.

Answer to comments of Reviewer#1

A) We would like to thank Reviewer#1 for the careful reading of this manuscript and the thoughtful comments that have improved the quality of this manuscript. Furthermore, below are those comments that need more clarification.

Comments to “Technical note: Expanding ensemble forecasts with generative AI: application to volcanic clouds” by Mingari et al.

This manuscript presents a powerful generative AI method Variational Auto Encoder (VAE) and its application in volcanic ash forecasts. It demonstrates that the VAE produces quality results comparable to those of the physic-based numerical model FALL3D. Additionally, the VAE can generate large ensembles without the high computing demands required by physic-based numerical models. This manuscript is well written, clearly structured, the underlying methodology is systematically explained, the results are convincing. This manuscript deserves to be published. Still, I have several concerns and some minor comments and suggestions.

Concerns:

1. This is a purely methodological manuscript. It investigates a method rather than using the method to answer scientific questions in volcanic studies. In my opinion, this manuscript aligns more to the scopes of the GMD journal than with that of the ACP journal. I am not sure how strict these Journal definitions are.

B) We thank the reviewer for raising this important point regarding the scope of the manuscript. We agree that the primary contribution of this work is methodological and that GMD could be a better option for this type of contribution. Originally, this work was planned as a two-paper study: the first paper would describe the methodology, while the second would focus on its application to data assimilation. With this in mind, we considered ACP a suitable option for the present work. Following the review, we decided to carry out a major revision to address the reviewer's concerns, including latent-space data assimilation and an application of the methodology to the 2018 Etna eruption. We also made this change because, in the original manuscript, the purpose of the methodology was perhaps not sufficiently clear, making it appear closer to a simple resampling strategy. We therefore believe that the revised manuscript is now better aligned with ACP, whose current aims and scope explicitly include atmospheric modelling and machine learning.

2. Additional discussions/descriptions are required:

- Discussions on the additional values this study brings to volcanic ash forecasts: It reads like that the only added values of this study is that the VAE can replace physic-based numerical model in real-time forecasting without computing

demands as required by the numerical models. However, can the proposed method offer more values to volcanic study, crises management, and short-term decision-making? Adding more specific discussions would enhance the importance of this work.

- A) This was partly a consequence of our initial two-paper plan, in which the first paper focused mainly on the methodology, without fully presenting its broader purpose. In the revised manuscript, we have reframed the work around the reduced-dimensional representation of the physical model provided by the VAE. The most straightforward application is sampling in the latent space to generate new model realizations at very low computational cost. More importantly, this reduced representation enables the application of data assimilation methods directly in latent space, with the potential to reduce forecasting errors. Therefore, although computational efficiency remains a real benefit, it is no longer presented as the main advantage of the approach. The revised manuscript shifts the focus from low-cost ensemble augmentation towards dimensional reduction and its potential for improving ensemble forecasting and data assimilation. This is also reflected in the new title, which makes clear that the methodology is intended as a complementary tool to FALL3D, rather than as a replacement for the physical model.

Modified title:

Integrating generative AI with physics-based simulations for probabilistic modelling of volcanic clouds

- Discussions on potential applications of this study: Potential applications of this study are mentioned briefly with short sentences in the manuscript (the last sentence in the abstract and two similar sentences in the introduction and conclusion). The current discusses are superficial and needed to be expanded with more details and to make them more convincing. Moreover, applications of this work are not limited to volcanic studies, it can be applied to a much broader field in atmospheric science.

- A) We agree with the reviewer that the potential applications of this study were only briefly discussed in the original manuscript. We also agree that the methodology is not limited to volcanic applications; rather, it can be applied to a much broader field within atmospheric science, such as dust storms and wildfires. In the revised manuscript, we have added a specific discussion section (Sect. 5.4) where we discuss the broader implications of the methodology and potential applications, while keeping these statements appropriately cautious.

Added in Sect. 5.4 - Discussion (line 561):

Although the present study is centred on volcanic ash applications, the proposed methodology could be naturally extended to other atmospheric modelling and forecasting problems involving the transport and dispersion of atmospheric components. Potential applications include air-quality

forecasting and, particularly, modelling of smoke transport from large wildfire events.

- Discussions on the forecasting lead time dependency: This manuscript addresses only 24-hours forecasts. Do only 24-hours forecasts matter for the crisis management? How about e.g. 48-hours forecasts? The authors do state in the manuscript that ‘The inclusion of time dependency is straightforward and can be done by simply expanding the dimension of the training dataset by adding successive timestamps.’. If so, an additional plot showing VAE performance on forecasting time dependency will be helpful. Meanwhile, it is well known that forecasting accuracy is highly dependent on the forecasting lead time. To certain forecasting lead time, forecasting model generate outputs which don’t have quality anymore and good training data is critical for machine learning applications. Will forecasting leading time affect the accuracy of VAE model, to what time extent can VAE be used for volcanic ash forecasting?

A) We thank the reviewer for raising this point. The choice of a single forecast time in this study is related to our focus on data assimilation. In a DA cycle, the analysis step corrects a prior ensemble using observations available at a specific time to produce an updated (posterior) ensemble. Therefore, for the latent-space DA application considered here, the VAE only needs to be trained for the analysis time being assimilated. In a typical operational workflow with two or four daily assimilation cycles, the procedure could consequently be applied at the corresponding analysis times, without requiring the time dimension to be included in each training dataset (although the training must be repeated for each assimilation cycle). On the other hand, applications explicitly requiring time dependency need a different treatment, and our previous statement that including time dependency is “straightforward” was therefore misleading and has been removed. For example, state-of-the-art atmospheric foundation models such as Aurora explicitly encode temporal information within their architecture. The appropriate treatment of time dependency is outside the scope of the present study and would likely require a major restructuring of the neural-network architecture.

Removed from Sect. 3.1: Training

~~The inclusion of time dependency is straightforward and can be done by simply expanding the dimension of the training dataset by adding successive timestamps.~~

Minor comments:

Abstract: missing quantitative statements for results. E.g., instead of saying ‘The statistical properties of the expanded ensemble are characterized in detail’, stating what statistical properties are characterized and what are the values of characterized statistical properties. The same for evaluation metrics.

- A) This suggestion has been addressed. We introduced a few metrics values and quantitative information in the updated abstract.

Added to abstract:

Two numerical experiments demonstrate that VAE-generated ensembles can reproduce key statistics of a larger reference ensemble and improve agreement with satellite retrievals. The VAE achieves up to a 1000-fold compression of the original fields into 16- and 32-dimensional latent spaces. The VAE-generated ensemble of 4096 members reduces RMSE by 4\% relative to the model prior, while latent-space data assimilation further reduces RMSE by 30\%.

Line 58: the definition of 'latent space' is not clear and precise. E.g, replace 'typically of lower dimensionality than the original data space,' to something like 'latent space is a lower-dimensional space of compressed and hidden representative of original data,'.

- A) This suggestion has been addressed.

Added to Sect. 1 - Introduction (line 55):

(..) Variational Autoencoders (...) learn a probabilistic mapping between a high-dimensional data space and a lower-dimensional latent space, providing a compact representation of the dominant features of the original data (...)

Line 76, 'FALL3D can produce an ensemble of simulations in a very efficient way by running a parallel job with multiple GPUs (Folch et al. 2022)'. The 'efficient' in this sentence is missing leading. If the physical-base model FALL3D is efficient enough, then the usage of VAE become less important.

- A) We thank the reviewer for this comment. We agree that FALL3D can generate ensembles efficiently when parallelized over multiple GPUs, as stated in the manuscript. However, this efficiency strongly depends on access to high-performance computing (HPC) resources. Many volcanological observatories worldwide, particularly in resource-limited settings, do not have access to such GPU capabilities. In these cases, generating large ensembles with physics-based models remains limited by the computational resources available. Therefore, the proposed AI-based approach provides a valuable alternative by enabling the generation of large ensembles at a significantly reduced computational cost (including a standard laptop), making it particularly relevant for operational settings with limited resources. This is further discussed in the Discussions section.

Added in Sect. 5.4 - Discussion (line 528):

Most importantly, the approach substantially reduces the computational resources required to generate large ensembles; for example, the simulations for the synthetic experiment on the LUMI supercomputer required at least 2,048 GPUs.

Line 85: replace 'for VAE training and evaluation' to 'for VAE training, testing and evaluation'.

A) This suggestion has been addressed.

Line 100: link or reference for this dataset 'EU-MODEX Tenerife 2025'?

A) This suggestion has been addressed.

Added in References:

European Commission: Field exercises for EUCPT and TAST "Cycle 12" – Lot 4: EU-MODEX Tenerife, Project brief, Union Civil Protection Knowledge Network,
<https://civil-protection-knowledge-network.europa.eu/projects/field-exercises-eucpt-and-tast-cycle-12>, 2025.

Line 105: the authors state 'the distinction between real and hypothetical eruption scenarios is not relevant here.' But it is interesting for readers to know the differences between real and hypothetical eruptions scenarios. Some extra short explanations might be great.

A) This point is addressed in the revised manuscript by including an additional experiment based on a real case: the 24 December 2018 Etna eruption, together with a comparison against satellite retrievals. In the original version, by stating that the distinction was not relevant, we meant that the modelling strategy used for the hypothetical eruption followed the same approach as that used for real eruptions. After all, the purpose of the exercise was to test the response to a potential eruption in the Canary Islands. Importantly, we selected this particular hypothetical case because its ensemble exhibits a large variability, with multiple qualitatively different plume solutions. This provided a useful test case to stress-test the ability of the VAE to reproduce a diverse range of volcanic-cloud scenarios.

Added to Sect. 4.2. Synthetic experiment (line 233):

Since our study focuses on generating ensembles from numerical simulations and approximating their statistical distribution, the distinction between real and hypothetical eruption scenarios does not affect the methodology or conclusions. This case was deliberately selected because the ensemble simulations produce qualitatively different plume transport patterns, providing a suitable test bed to inspect the ability of the VAE to reproduce the variability and statistical characteristics of the simulated ensemble.

Table 1 'Number of grid points: 120×100×25': later on, in Section 3.1 and Table 4, grid size is 120×100 (2D grid). Is the '25' the time dimension which has been removed from the manuscript?

A) No, it corresponds to the number of vertical levels (Table 1 has been updated to make this clear). The grid size 120x100x25 refers to the three-dimensional FALL3D computational grid, with dimensions along longitude, latitude, and vertical levels. FALL3D simulates the three-dimensional ash concentration for different particle classes. For the VAE training, however, we use volcanic ash column mass, obtained by vertically integrating the concentration, resulting in the 120x100 two-dimensional fields. Extending the methodology to the full three-dimensional concentration field is one of the challenges identified for future work and is discussed in Sect. 5.4.

Table 1 ‘Ensemble size 64, 256 and 2048’: this information is duplicated in Table 3. Either remove this information from Table 1, or remove Table 3 but add information of ‘Training’, ‘Validation’ and ‘Test’ to Table 1.

A) Suggestion accepted. Table 3 was removed and the ensemble size information is reported only in Table 1.

Line 113-119 description to Table 2: missing arguments for the chosen parameter sampling ranges. Why $\pm 30\%$ and $\pm 25\%$ are chosen for these parameter perturbations?

A) There is no universally accepted criterion for defining these perturbation ranges. Since this configuration corresponds to a controlled experiment, the perturbation ranges were chosen to provide sufficient variability within the ensemble and represent different plausible volcanic scenarios, allowing us to test the ability of the VAE to capture this variability. At the same time, the perturbations were selected to represent plausible variations in eruption source conditions over a typical day of activity, based on operational experience and previous studies of eruptive variability.

Line 120-124: are the usage of ‘simulation A’, ‘simulation B’ and ‘simulation C’ necessary? Using ‘training’, ‘validation’ and ‘test’ seem clear enough for me.

A) Suggestion accepted. We removed the unnecessary notation from Sect. 4.2. (Synthetic experiment) when referring to different simulations, and now use the clearer terms “training”, “validation” and “test” instead.

Line 130: the sentence of citation ‘(Kingma and Welling, 2013)’. This sentence explains only half of the VAE, it is better to move the citation to one sentence after.

A) Suggestion addressed.

Line 131: remove ‘vectors z ’ here. Vector z is only clear afterwards when Fig.1 is explained.

A) Suggestion addressed.

Figure 1: merge the two boxes for encoder, and the two boxes for decoder. These two boxes supposed to mean two hidden layers, and it is not necessary that encoder and decoder always have two hidden layers. It is clear enough to state in the text that the encoder and decoder used in this study have two hidden layers.

A) Actually, the decoder and encoder in our work are each composed of three layers, resulting in a total of six Conv2d hidden layers. We agree that the figure represents a very specific architecture, while VAEs generally don't require two layers. However, we prefer to retain the current representation because the two layers are shown deliberately to illustrate the progressive dimensionality reduction (expansion) through the encoder (decoder) without overloading the figure. We modified the figure caption to make this clear.

Added to caption of Fig. 1:

For illustrative purposes, the encoder and decoder are shown with two hidden layers, representing the progressive dimensionality reduction and expansion. In this work, we used three (synthetic experiment) and four (Etna experiment) hidden layers for the encoder and decoder.

Figure 1 Caption: do 'two-dimensional convolutional layers' mean 'Conv2D' or two convolutional layers?

A) It's Conv2D. Now clarified in the caption.

Updated to caption of Fig. 1:

The basic scheme of a VAE composed of two-dimensional convolutional layers (Conv2D).

Line 137: for clarity, it is better to add '(the first term)' after 'the fidelity of the data reconstruction', and add '(the second term)' after '... a prior distribution $p(z)$ '.

A) Suggestion addressed.

Line 147: it might be clearer to replace 'measures the difference...' to 'measures the distance...'

A) Suggestion addressed

Line 150: what is the reason to write $p(z)=$, instead of $p(z) = N(0,1)$? More specifically, what is special about this ? It has been also used the line above (Line 149).

- A) It reflects the multidimensional nature of the latent space: instead of using the univariate notation $N(0,1)$, we use this notation to explicitly indicate a multivariate standard normal distribution with zero mean and identity covariance matrix.

Figure 2 Caption: there is a typo. Replace '... of the latent space (a).' to '... of the latent space (b).'

- A) Corrected.

Line 239 and 240: according to the descriptions, boxplots are calculated based on these 20×2048 ensembles. It is still not clear for me how to deal with 2D grid points, do you calculate mean of 120×100 grid points first, then apply statistic metrics to ensembles?

- A) For each VAE-generated ensemble, we compute the ensemble mean at every grid point, resulting in a field (x) on a 120×100 grid, such as the one shown in Fig. 5. Similarly, the ensemble mean of the test dataset is computed to obtain the reference field ($\hat{x}$). The RMSE is then computed over the 120×100 grid points ($N=12,000$):

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \hat{x}_i)^2}$$

This process is repeated for each of the 20 newly generated ensembles, resulting in 20 RMSE values that are then used to construct the boxplots. The full code used in this work is available in the repositories listed in the supplementary material.

References:

Mingari, L.: Supplementary material for egusphere-2025-6097: Synthetic experiment, <https://doi.org/10.5281/zenodo.22759580>, 2026a.

Line 242-261 regarding to the selection of $K=16$ and $\beta=6$ as optimal hyperparameters: is this decision made purely based on a minimal RMSE? As the authors discuss in this section, higher latent dimensions lead to less stable training and less structured latent space, then why not select $K = 6$ as optimal K ? As we can see from Figure 4a, $K=6$ and $K=16$ result in very similar RMSE values. What is the reason to prefer $K=16$ to $K=6$?

- A) The selection of $K=16$ and $\beta=6$ was not based solely on minimal RMSE. While Figure 4a shows that $K=6$ and $K=16$ achieve similar RMSE values, other considerations guided our choice:
- With $K=6$, the latent space is highly compressed, which limits the model's ability to capture complex correlations and variability in the data. By using $K=16$, the latent space has more capacity to represent underlying factors, leading to a more structured and informative latent representation.

- Although large latent spaces can be less stable if excessively large, we found that $K=16$ balances capacity and stability, allowing the model to learn meaningful latent features without sacrificing convergence.
- A higher-dimensional latent space allows the decoder to receive a richer, more informative latent code, improving the diversity and interpretability of generated ensembles, even if the RMSE of the mean is similar to that of $K=6$.

In summary, $K=16$ was preferred because it provides a better trade-off between latent space expressiveness, interpretability, and training stability, not just RMSE minimization.

Line 262: section 4.2 has a title 'Ensemble mean and spread', but this section only talks about mean, the results of spread are presented in section 4.3 (Figure 8 and its description texts). Is it better to change the title of Section 4.2 to 'Ensemble mean'?

- A) We decided to keep the title "Ensemble mean and spread" because this section also discusses the results reported in Table 5 (Table 3 in the revised manuscript), which includes evaluation metrics for both the mean and spread.

Updated text in Sect. 5.2.1: Ensemble mean and spread (line 394):
A summary of the metrics for the ensemble mean and spread is shown in Table 3.

Line 276: the sentence 'We can conclude that VAE is properly capturing the first and second moments of the distribution', what do 'first and second moments' mean here and how are they linked to ensemble means?

- A) The first and second moments of a distribution correspond to its mean and variance, respectively. In our analysis, these are estimated from the ensemble mean and spread (standard deviation), with the latter directly related to the variance.

Line 283: to be more precise, it is better to include ensemble member size in the statement. Maybe modify 'demonstrating that the VAE is capable of ...' to something like 'demonstrating that with an ensemble member of 5000 VAE is capable of ...'. Because as shown in Figure 9, with small sample member less than 256, VAE can't reproduce the results of training data (as shown in Figure 9).

- A) Agreed. This has been clarified by explicitly stating the ensemble size.

Updated text in Sect. 5.2.2 Probability distributions (line 409):
(...) demonstrating that a 5000-member VAE-generated ensemble can produce accurate probability distributions as well as realistic samples.

Figure 7: add a legend for contour lines of 2%, 25%, 50%, 75% and 98%. Or use more distinct colors for these specified probability values and state line colors in the figure caption, e.g. 2% (purple) etc.

- A) The colours and line types were modified to make it easier to identify each probability contour in the caption.

Updated caption of Fig. 6:

Exceedance probabilities for three ash column mass thresholds (8, 16 and 24 g m⁻² from top to bottom). The probability contours indicate probability levels of 2% (blue), 25% (green), 50% (dashed black), 75% (orange), and 98% (magenta). Results for the training dataset (left column) and a VAE-generated ensemble with 5,000 samples (central column) are compared with the test dataset (right column).

Line 313: regarding to 'Future work should investigate more advanced neural network architectures to better address this limitation.' , the question is if this limitation is imbedded in training data itself (incapacity of model). If the signal is not in the training data, it doesn't matter how advance the used neural network architectures are, the situation will not be improved.

- A) We agree that part of this limitation might originate from the training dataset itself, although we have not explored this aspect in detail. However, later experiments indicate that the high-frequency artifacts are also related to the VAE, particularly to the decoder architecture. Specifically, we found that the transposed-convolution upsampling used in the original manuscript introduced checkerboard patterns, which can be reduced by using bilinear interpolation before the convolutional operation in the decoder. Therefore, the loss of high-frequency content observed in the VAE-generated samples is at least partially related to the model's limited ability to reproduce these fine-scale structures. This is further discussed in Sect. 5.2.3 and shown in the updated Fig. 9.

Added to Sect. 5.2.3 - Isotropic power spectrum (line 437):

A new neural network was trained using a decoder based on bilinear interpolation for upsampling instead of transposed convolution layers, in order to investigate whether this alternative decoder architecture leads to stronger low-pass filtering (blue crosses in Fig. 9). Although deviations from the test spectrum persist for $k > 0.33 \sim \Delta x^{-1}$, the high-wavenumber spectral power is noticeably reduced, indicating that the bilinear-interpolation decoder provides more effective suppression of small-scale noise than the original transposed-convolution decoder.

Line 335-336: add quantitative description of typical model error, by extending the sentence 'This is negligible compared to ... in real applications.' with ', which is in the range of xxx-yyy.'

- A) This is addressed explicitly in the revised manuscript, where the agreement of both the VAE-generated ensemble and the training ensemble with the observations is quantitatively evaluated in a new numerical experiment (Etna experiment). The comparison shows similar performance, with the VAE-generated ensemble outperforming the training ensemble in terms of RMSE, RMSLE, and Pearson correlation coefficient. See Sect. 5.3.1 for further details.

Line 340: specify ‘dramatically reduce’

- A) A new experiment has been added in the revised manuscript to provide a preliminary quantification of the data assimilation effect. See Sect. 5.3.2, “Data assimilation experiments”, where the latent-space data assimilation results are presented. The term “dramatically” was removed from the last sentence of the conclusions, which was reformulated in more cautious terms.

Updated last sentence of Conclusions (line 591):

In summary, this work represents a first proof of concept based on a relatively simple neural architecture and an elementary data assimilation methodology. The results demonstrate the potential of the formulation, while highlighting the need for further research to assess its robustness and applicability across different eruption scenarios. Future work should extend the methodology to three-dimensional fields, explore more advanced generative models and assimilation approaches, and evaluate their performance against state-of-the-art DA methods and independent observations, ultimately assessing their potential implications for operational forecasting.

Answer to comments of Reviewer#2

A) We would like to thank Reviewer#2 for the careful reading of this manuscript and the thoughtful comments that have improved the quality of this manuscript. Furthermore, below are those comments that need more clarification.

This technical note uses a Variational Autoencoder to expand ensemble forecasts of volcanic ash clouds generated by the FALL3D model. I think the topic is interesting and potentially useful for operational volcanic-cloud forecasting, especially because large physics-based ensembles are computationally expensive. The manuscript is generally clear, and the authors have made a reasonable first attempt to evaluate the generated ensemble against a larger numerical ensemble. In my view, the weakest point is that it does not yet clearly demonstrate whether the proposed VAE method provides real added value beyond a statistical interpolation or smoothing of one specific numerical ensemble. The study is based on one hypothetical eruption case, one model setup, and one forecast time. Therefore, I think the current manuscript is more like a proof-of-concept than a method that has been convincingly demonstrated for broader volcanic-cloud forecasting. I suggest a major revision before it can be considered for publication in ACP.

Major:

1. I understand that this is a technical note, and a simple case study is acceptable as a first step. However, the current manuscript sometimes gives the impression that the method is already generally useful for operational ensemble forecasting. I do not think this has been demonstrated. A natural question is: if the volcano, eruption height, emission rate, wind situation, or forecast time changes, can the same trained VAE still be used? Or does a new VAE need to be trained for each new forecast case? If the latter is true, the operational value of the method is more limited. I suggest that the authors either add another case or clearly state that the current work only demonstrates feasibility for a single controlled case.

A) General comment: We thank the reviewer for this insightful comment, which led us to substantially revise the original manuscript. Originally, this work was planned as a two-paper study: the first paper would describe the methodology, while the second would focus on one of its applications, i.e. data assimilation. As a result, the purpose of the original manuscript was not sufficiently clear, and the methodology could indeed appear closer to an overcomplicated resampling technique. This is far from true, however...

We have reframed the focus of the manuscript during the revision. Although the reduction in computational cost is an important benefit, we now emphasize the reduced-dimensional representation learned by the

VAE as the primary advantage. This provides a basis for developing new approaches to probabilistic modelling of volcanic clouds and, potentially, other atmospheric transport applications. Data assimilation is one such application, with the potential to improve forecasting skill by correcting the ensemble in the reduced-dimensional latent space. To make this clearer, we have reformulated the title of the manuscript:

Modified title:

Integrating generative AI with physics-based simulations for probabilistic modelling of volcanic clouds

- A) New experiment: During the revision, and following the reviewer's suggestion, we decided to carry out a major revision that substantially broadens the scope of the study. In particular, we added a new experiment based on the 24 December 2018 Etna eruption, including a comparison with satellite observations and the testing of two latent-space data assimilation methods. These additions provide a first assessment of the methodology beyond the original controlled experiment, while also demonstrating how the reduced-dimensional representation can be exploited for applications beyond ensemble generation.
- A) Transition to operations: This work is basic research exploring the potential benefits of dimensionality reduction for improving probabilistic modelling of volcanic clouds. It remains a proof of concept, potentially opening promising new research lines, rather than a study of transition to operations. We have therefore removed references that could mislead the reader into interpreting the methodology as being ready for operational applications. Operational forecasting is now mentioned only as a possible future application at the end of the Discussion and Conclusions sections.

Added in Sect. 5.4 - Discussion (line 564):

Future research is needed to assess whether the methodology can be extended to the full three-dimensional concentration field and multiple particle classes, which will require adapting the neural network architecture to the increased dimensionality of the problem. This extension would be an important step for enabling multiple data assimilation cycles and, ultimately, supporting operational forecasting.

Added in Sect. 6 - Conclusions (line 591):

In summary, this work represents a first proof of concept based on a relatively simple neural architecture and an elementary data assimilation methodology. The results demonstrate the potential of the formulation, while highlighting the need for further research to assess its robustness and applicability across different eruption scenarios. Future work should

extend the methodology to three-dimensional fields, explore more advanced generative models and assimilation approaches, and evaluate their performance against state-of-the-art DA methods and independent observations, ultimately assessing their potential implications for operational forecasting.

A) Re-training: The proposed methodology is not a foundational model intended to be applied unchanged to arbitrary volcanic or meteorological conditions. If the conditions of the problem change (e.g., volcano, meteorological situation, eruption characteristics, ...), a new physics-based ensemble simulation would be required and the VAE would need to be re-trained. The methodology should therefore not be viewed as a replacement for the physics-based numerical model, but rather as a complementary tool that learns from the ensemble generated by FALL3D. We believe, however, that the computational cost of this re-training can still be advantageous in an operational context. As an illustrative example, volcanic ash advisory centres typically provide forecasts at +6, +12, and +18 h. Under this approach, a physics-based ensemble should be generated and three VAE configurations trained for these forecast times. Under the same experimental conditions as the Etna experiment (see Sect. 4.3), generating a 128-member FALL3D ensemble on a supercomputer requires approximately 2.25 GPU-hours (18-h forecast), whereas training each VAE requires approximately $6\text{E-}3$ GPU-hours on a laptop equipped with a GPU. Thus, training three VAE configurations would require only approximately 0.017 GPU-hours in total. The computational cost of inference is negligible compared with training, allowing thousands of additional VAE-generated ensemble members to be produced at essentially no additional computational cost. In contrast, increasing the size of the FALL3D ensemble scales the computational cost and the associated requirements in terms of GPU-hours and computational nodes. Let's say we want to expand the training ensemble to 1,000 simulations. This would require around 18 GPU-hours with FALL3D, i.e. roughly a 1,000x increase compared with the GPU time required by the VAEs. Therefore, although re-training is required when the underlying conditions change, the ability to generate very large ensembles from a relatively small physics-based ensemble can still provide a substantial computational advantage, particularly for observatories and operational centers with limited access to HPC resources. This is also true for data assimilation applications. Operational centres normally perform two or four assimilation cycles per day. Similarly, latent-space data assimilation would require the VAE to be trained two or four times per day, depending on the number of assimilation cycles. This is perfectly feasible!

A) Size of the training dataset: But the real important question here is: why is the training so fast? And this question is directly related to the size of the training dataset. The relatively small size of the training dataset is an

important point here and is directly related to the short training time. In this work, the training dataset consists of only a few hundred simulations. In our opinion, one of the interesting results of the study is that this amount of data is sufficient for the VAE to learn a useful representation of the solution space. When referring to the “operational value of the method”, we understand that the reviewer may have in mind applications requiring huge training datasets and long training sessions. This is not the case here. We are not training a neural network to distinguish images from highly heterogeneous sources (animals, human faces, objects, etc.); rather, we are learning a low-dimensional manifold representing the solutions of a PDE solved by FALL3D. These solutions are continuous and well behaved, and therefore form a much more structured solution space. Our hypothesis is that this structure is one of the reasons why the VAE can be trained from a relatively small ensemble of simulations and subsequently used to efficiently augment it. We showed this is true for two use cases; whether this remains true for substantially different eruption scenarios is something that needs to be investigated in future works.

2. The authors show that the VAE-generated ensemble can approximate the 2048-member numerical ensemble, but it is unclear whether this improvement really comes from the VAE. For example, would a bootstrap method, kernel density estimation, PCA-based sampling, or another simple statistical resampling method produce similar results? This is a basic but important comparison. Without such a baseline, it is difficult to know whether the VAE is actually necessary, or whether it mainly smooths and interpolates the information already contained in the 256-member training ensemble. I suggest adding at least one simple baseline method.

A) We thank the reviewer for this valuable suggestion. Following this comment, we decided to carry out a major restructuring of the original manuscript. We realised that, in its original form, the message of the paper could be interpreted as “*this is just an overcomplicated resampling methodology*”. However, this is not the purpose of the proposed methodology, and we explain below why we do not consider a comparison with simple resampling strategies to be appropriate in this context.

First, we have now clearly emphasised that the focus of the work is on generative AI. The impressive ability of these models to generate entirely new samples is nowadays widely known. We investigate whether a generative model can learn the solution space represented by a physics-based model and generate new samples that are statistically consistent with that model. This is very different from simple resampling. The appropriate way to evaluate this ability is therefore to compare the generated samples against a large reference ensemble of independent numerical simulations. In addition, the comparisons proposed by the reviewer are not directly relevant to the revised version of the manuscript, since it is now clear that our purpose goes beyond ensemble

augmentation. Our objective is to take advantage of the dimensionality reduction of the original problem provided by the VAE. To do this, we first need to characterise the VAE and demonstrate that the generative model is able to produce new samples that are coherent with the solution space represented by the physics-based model. However, we are not looking for alternative ensemble expansion techniques, since ensemble augmentation is not the main objective of the proposed methodology.

PCA-based sampling: The VAE encoder learns a non-linear transformation between the physical space and the latent space. In this way, the decoder is able to transform samples from a standard normal distribution in the latent space into a high-dimensional distribution in the original physical space. This distribution is highly skewed and, in some cases, resembles gamma distributions (Mingari et al. 2023). In contrast, with PCA, samples are drawn in the reduced space and mapped back to the physical space through a linear transformation. Therefore, a Gaussian distribution in the reduced space would remain Gaussian in the physical space. The non-linear transformation learned by the VAE is a key element for representing the non-Gaussian distribution of the modelled plume. Atmospheric dynamics are highly non-linear and cannot, in general, be represented through linear transformations.

Another way of looking at this is that the VAE is composed of multiple layers, including convolutional and linear layers, which allow it to learn this non-linear transformation. The reviewer's suggestion would be somewhat equivalent to asking us to take the VAE and eliminate all of these layers, retaining only a single linear layer. Such a simplification would obviously deteriorate the performance of the model and, more importantly, would not provide particularly useful information about the methodology. A more scientifically relevant sensitivity analysis would instead investigate the effect of the number of layers and other parameters of the neural-network architecture.

Bootstrap method: Bootstrap resampling is inherently restricted to combinations of already existing realizations and therefore cannot generate new physically plausible scenarios beyond those represented in the training ensemble.

Kernel density estimation: Kernel density estimation (KDE), although non-parametric, suffers from the curse of dimensionality and does not explicitly learn a compact latent representation of the ensemble. In contrast, the VAE learns a continuous nonlinear latent manifold that provides a probabilistic representation of the underlying data distribution. This allows the generation of entirely new realizations while preserving complex spatial correlations and potentially multimodal structures. This point is particularly relevant because volcanic ash hazard maps are high-dimensional spatial fields. In this context, the VAE does not simply interpolate or smooth the information contained in the 256-member

training ensemble. Instead, it learns a compact nonlinear representation of these fields, enabling the generation of new physically plausible scenarios that approximate the statistical characteristics of much larger ensembles.

References:

Mingari, L., Costa, A., Macedonio, G., and Folch, A.: Reconstructing tephra fall deposits via ensemble-based data assimilation techniques, *Geosci. Model Dev.*, 16, 3459–3478, <https://doi.org/10.5194/gmd-16-3459-2023>, 2023.

3. The small-scale error shown in the spectral analysis should be discussed more seriously. Figure 10 shows that the VAE can reproduce large-scale structures, but it fails at smaller spatial scales and introduces artificial high-wavenumber noise. This may not be a small technical detail. In volcanic ash forecasting, plume edges, local gradients, and threshold exceedance areas can be important for aviation warnings. If the VAE changes or smooths these small-scale structures, the exceedance probabilities and hazard boundaries may also be affected. I suggest that the authors discuss whether this limitation matters for practical forecasting.

A) We thank the reviewer for highlighting this important point. We agree that the discrepancies observed at small spatial scales in the spectral analysis deserve careful consideration, particularly in the context of volcanic ash forecasting. The VAE reproduces the dominant large-scale spatial structures of the FALL3D ensemble, but the original decoder also introduces artificial high-wavenumber noise.

To investigate this limitation, we have added a new experiment using an alternative decoder architecture. The results show that the high-frequency noise can be substantially reduced through modifications to the neural network architecture. In addition, we have computed the contribution of the high-frequency spectral power to quantify the magnitude of this discrepancy. Although the high-wavenumber contribution represents only a small fraction of the total spectral power, we agree that its potential implications should be considered carefully, since small-scale gradients and plume boundaries may influence threshold exceedance areas and hazard boundaries relevant to aviation warnings.

We have therefore expanded the discussion in the revised manuscript to explicitly acknowledge the limitations of the VAE at small spatial scales and their potential implications for volcanic ash forecasting. At the same time, the new results indicate that this limitation can be mitigated through appropriate choices of neural network architecture. In addition, the impact of this noise on the results presented in this work is not significant, partly because it is averaged out in the ensemble statistics. Then, we remain cautious about the relevance of these errors, as their potential implications have not been directly assessed in the present study.

Added to Sect. 5.2.3 - Isotropic power spectrum (line 437):

A new neural network was trained using a decoder based on bilinear interpolation for upsampling instead of transposed convolution layers, in order to investigate whether this alternative decoder architecture leads to stronger low-pass filtering (blue crosses in Fig. 9). Although deviations from the test spectrum persist for $k > 0.33 \Delta x^{-1}$, the high-wavenumber spectral power is noticeably reduced, indicating that the bilinear-interpolation decoder provides more effective suppression of small-scale noise than the original transposed-convolution decoder.

To quantify the differences at small spatial scales, we computed the fraction of total spectral power associated with wavenumbers $k > 0.33 \Delta x^{-1}$. This fraction was 0.003% for the test dataset, 0.020% for the VAE with the transposed-convolution decoder, and 0.009% for the VAE with the bilinear-interpolation decoder. In other words, the bilinear-interpolation decoder reduced the excess power by approximately 55% relative to the transposed-convolution decoder.

Although the high-wavenumber power contribution represents only a small fraction of the total spectral power, its relative excess may still affect the representation of fine-scale plume structures. In particular, errors near plume boundaries and in local concentration gradients may affect the spatial extent of areas defined by threshold exceedances, which are relevant to volcanic ash forecasting and aviation hazard assessment. Assessing this potential impact requires further investigation.

4. The computational-cost argument needs to be clearer. The authors emphasize that VAE samples can be generated very cheaply, but the full workflow also requires producing the 256-member training ensemble and training the neural network. If this has to be done for every forecast case, the computational advantage may be smaller than suggested. I suggest that the authors provide a clearer cost comparison: running 256 FALL3D members plus VAE training and generation versus directly running a larger FALL3D ensemble. The authors should also clarify whether the trained VAE is expected to be reused across different forecasts.

A) We added computational cost metrics in the Discussion section (Sect. 5.4). The metrics are based on the new Etna experiment, which is computationally more demanding and more representative of an operational workflow. For this case the 3D computational grid was expanded to 256x128x40 (longitude, latitude, vertical levels). We summarise the results in terms of GPU-hours for 8-h forecasts and 128 ensembles members simulated. Note, however, that (i) the computational cost of the FALL3D simulations scales with both the total simulated time and the ensemble size and (ii) FALL3D requires a supercomputer with at least one GPU per ensemble member. For example, increasing the FALL3D ensemble from 128 to 1,000 members would require approximately 8 GPU-hours under the same configuration, whereas the additional cost of

generating the corresponding VAE ensemble would remain negligible after the initial 0.006 GPU-hours required for training.

FALL3D simulations: 128 ensemble members, run on the accelerated partition of MareNostrum5 using 128 NVIDIA Hopper GPUs.

Computation cost: ~1 GPU-hour

VAE training: training performed on a laptop using a single RTX 5000 Ada Generation NVIDIA GPU.

Computation cost: ~0.006 GPU-hours

VAE inference: Negligible compared with the training cost, even when generating ensembles of up to 5,000 members on a laptop CPU. We did not test larger ensembles.

- A) Finally, the VAE needs to be trained for every new forecast time. For example, if a volcanic ash advisory centre (VAAC) performs four daily assimilation cycles, four training sessions would be required. However, given the relatively low computational cost of training, this overhead is not expected to be critical for the proposed application.

Added to section 5.4 - Discussion (line 520):

It is interesting to consider whether this methodology can be computationally efficient, accounting for both training and inference, when generating ensembles much larger than the training ensemble. However, the answer is not simple as it depends on the available computational resources, the cost of the physics-based model, the total simulation time, and the ensemble size, among other factors. As an illustrative example, let's consider the Etna experiment, which is the most computationally demanding case presented here. The 128 FALL3D simulations were run on the accelerated partition of MareNostrum5 using 128 NVIDIA Hopper GPUs, requiring approximately 1 GPU-hour in total (i.e., aggregated across all GPUs). In contrast, VAE training was performed on a laptop using a single RTX 5000 Ada Generation NVIDIA GPU and required only 5.83×10^{-3} GPU-hours. In this experiment, the forecast period was limited to 8 h; extending the forecast period would further increase the relative cost of the numerical simulations. Most importantly, the approach substantially reduces the computational resources required to generate large ensembles; for example, the simulations for the synthetic experiment on the LUMI supercomputer required at least 2,048 GPUs. Of course, these examples do not allow us to draw general conclusions about the relative efficiency of augmenting an ensemble with a VAE versus explicitly simulating larger ensembles, as the computational cost depends on the type of model and the available computational resources.

5. The 2048-member FALL3D ensemble is treated as the target distribution, but it is still a model-generated reference, not the true atmospheric distribution. I think

the manuscript should be more careful when using terms such as “true distribution” or “actual distribution”. The VAE is evaluated against a larger FALL3D ensemble, not against real observations. This distinction should be made clearer, especially when discussing possible operational applications.

- A) Suggestion addressed. All references to “true distribution” or “actual distribution” have been replaced by “reference distribution”. We have also significantly reduced references to operational applications, emphasising that this work represents a first step towards exploring the complementary use of generative AI and numerical simulations.

Minor:

1. The manuscript repeatedly uses expressions such as “physically consistent” or “physics-grounded samples”. I understand what the authors mean: the VAE is trained using outputs from a physical model. However, the generated samples themselves are not produced by solving the governing equations. Therefore, I think this wording is a little too strong. The generated fields may be statistically similar to FALL3D outputs, but this does not necessarily mean that they satisfy physical constraints such as mass conservation or transport consistency. I suggest either adding some basic physical checks, or using more cautious wording, such as “statistically consistent with the FALL3D ensemble”.

- A) Suggestion addressed. Expressions like “physically consistent” or “physics-grounded samples” were removed. For example:

Added to Sect. 6 - Conclusions (line 579):

Even when no new physical knowledge is introduced by a VAE-augmented ensemble, it provides a good statistical approximation of the reference distribution and generates fields that are statistically consistent with the underlying physical model, particularly at large spatial scales.

2. The title and abstract should make it clearer that this is a proof-of-concept study based on one volcanic-cloud simulation setup.

- A) The title and abstract have been completely reformulated to better represent the proposed methodology and its proof-of-concept nature.

Modified title:

Integrating generative AI with physics-based simulations for probabilistic modelling of volcanic clouds

Added to abstract:

As a proof of concept, the data assimilation problem is formulated in two dimensions.

3. Some claims in the conclusion are too strong. For example, the statement that the method can be extremely useful for ensemble expansion and operational frameworks should be softened unless more cases are tested.

A) Suggestion addressed. The Conclusions have been completely reformulated to better represent the purpose and limitations of the methodology, with only a single reference to its potential implications for operational forecasting.

Added in Sect. 6 - Conclusions (line 591):

In summary, this work represents a first proof of concept based on a relatively simple neural architecture and an elementary data assimilation methodology. The results demonstrate the potential of the formulation, while highlighting the need for further research to assess its robustness and applicability across different eruption scenarios. Future work should extend the methodology to three-dimensional fields, explore more advanced generative models and assimilation approaches, and evaluate their performance against state-of-the-art DA methods and independent observations, ultimately assessing their potential implications for operational forecasting.

4. The authors mention future applications in latent-space data assimilation. This is interesting, but it is not tested in this manuscript. I suggest keeping this part shorter and more cautious.

A) Two latent-space data assimilation methods have been added to the revised manuscript. This significantly clarifies the objective of the work and demonstrates a potential application of the methodology to the forecasting of volcanic clouds. See Sect. 4.3 - Etna experiment.

5. It would be useful to explain more clearly why the authors selected the thresholds used in the exceedance probability analysis. Are these thresholds operationally meaningful, or just chosen for demonstration?

A) The three thresholds were selected for demonstration to span the range of values represented in the scales used for the maps (5-60 g/m², see Fig. 4), using low (8), medium (16) and high (24) values. Volcanic ash column-mass concentrations can span several orders of magnitude depending on the eruption, so there are no universal criteria for this. We selected values relevant for this specific experiment. For comparison with satellite retrievals, however, the thresholds considered here are well above the typical detection threshold for satellite-retrieved column mass (~0.1 g/m²). Therefore, the selected values are also relevant for comparing the modelled and satellite-retrieved column mass.

6. The manuscript should state more explicitly whether the VAE was trained only on ash column mass at +24 h. If so, the limitation of ignoring temporal evolution should be emphasized.

- A) We thank the reviewer for this comment. We now clearly specify in Section 4.4 that the VAE is trained using ash column mass fields at a single forecast time (+24 h):

Added to Sect. 4.4 - VAE configuration (line 278):

Training was performed using the simulated ensembles of two-dimensional volcanic ash column-mass fields at a single forecast time.

- A) We emphasize that this is a deliberate design choice rather than a limitation of the methodology. Including temporal dependence would require more complex neural network architectures and additional encoding of the forecast lead time, which would introduce unnecessary complexity for the data assimilation framework considered here. Our approach is conceived as complementary to the physics-based numerical model rather than as a replacement for it. The numerical model provides detailed temporal evolution of volcanic ash clouds, while the VAE provides a more robust statistical representation of the ensemble at the required forecast times.

Added to Sect. 5.4 - Discussion (line 511):

In this first work, the VAE is trained at a single forecast time to simplify the neural network architecture and avoid the need for additional layers for time encoding. This design choice reflects the complementary role of the VAE and the physics-based model: while the numerical model provides detailed temporal evolution of volcanic ash clouds, the VAE provides a robust statistical representation of the ensemble at the forecast times required for data assimilation or ensemble augmentation. By focusing on a single forecast time, the VAE can exploit the redundancy implicitly present in the ensemble of physics-based simulations to learn a low-dimensional manifold, while retaining a relatively simple and robust neural network architecture.