**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.

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.

    - 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.

    - Discussions on the forecasting lead time dependency: This manuscript addresses only 24-hours forecasts. Do only 24-hours forecasts matter for the crise 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 expending 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 depend 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?

**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.

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,’.

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.

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

Line 100: link or reference for this dataset ‘EU-MODEX Tenerife 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.

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?

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.

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

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.

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.

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

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.

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

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)’.

Line 147: it might be clearer to replace ‘measures the difference…’ to ‘measures the distance…’

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

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

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?

Line 242-261 regarding to the selection of K=16 and β=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?

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’?

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?

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).

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.

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.

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.’

Line 340: specify ‘dramatically reduce’