https://doi.org/10.5194/egusphere-2025-6097
Preprint. Discussion started: 30 December 2025

Technical note: Expanding ensemble forecasts with generative AI: application to volcanic clouds

Leonardo Mingari[1], Arnau Folch[1], Heribert Pascual[1], and Manuel Titos[2]

[1]Geosciences Barcelona (GEO3BCN-CSIC), Barcelona, Spain
[2]University of Granada, Granada, Spain

Correspondence: Leonardo Mingari (lmingari@csic.es)

Abstract. Ensemble-based modelling of the atmospheric dispersal of volcanic clouds enables more realistic forecasts by explicitly accounting for uncertainties in eruption source parameters, meteorological data, and systematic errors in transport models. Many ensemble applications, including quantification of forecast uncertainties, data assimilation, or probabilistic hazard assessments, require a large number of members to mitigate sampling errors and to properly capture probability distributions. However, running large ensembles with Volcanic Ash Transport and Dispersal (VATD) models can be computationally demanding, even for high-performance computing clusters. As a result, operational forecasting is typically restricted to smaller ensembles in order to fit time-to-solution requirements. In contrast, generative AI models can produce large volumes of physically-consistent data samples with minimal computational cost. In this work, a convolutional Variational AutoEncoder (VAE) is trained on an ensemble of 256 forecasts simulated with the FALL3D model and subsequently used to generate larger ensembles, effectively augmenting physics-based ensemble modelling capacity. Ensembles with up to 8192 members were generated nearly instantaneously using the trained neural network, with no reliance on HPC resources. The statistical properties of the expanded ensembles are characterised in detail, and the VAE performance is evaluated against a test dataset composed of 2048 numerical simulations. The VAE-generated ensembles closely approximate the actual (target) probability distribution as well as key sample statistics, such as ensemble mean and spread, with minimal degradation in the evaluation metrics. Finally, we discuss possible future applications of this work, including latent space data assimilation via deep learning.

1 Introduction

Operational forecasts of volcanic ash and SO_2 clouds rely on physics-based numerical simulations that yield airborne concentration fields and the corresponding aviation-hazard products. In order to produce realistic forecasts, Volcanic Ash Transport and Dispersal (VATD) models must describe several complex processes, including plume micro-scale dynamics, atmospheric long-range transport, particle sedimentation and deposition or the injection of volcanic particles and aerosols into the atmosphere. However, the complexity of the underlying physical processes, the broad range of spatial and temporal scales involved, and the uncertainties in the quantification of the emissions and meteorological fields make the modelling of volcanic clouds particularly challenging and the resulting forecasts prone to errors. Specifically, the so-called Eruption Source Parameters (ESPs), such as those required to estimate the vertical distribution of mass along the eruptive column, mass emission rates

https://doi.org/10.5194/egusphere-2025-6097
Preprint. Discussion started: 30 December 2025

or particle size distributions, are recognised to be first-order contributors to model errors (Costa et al., 2016; Poulidis and Iguchi, 2021). In fact, the determination of the ESPs is one of the major challenges in volcanic plume modelling due to limited real-time observations, complex eruptive dynamics and the rapid variability in emission patterns. Ensemble modelling is a popular approach to deal with the uncertainties associated with model parameters, the spatial distribution and time dependency of emissions or the underpinning meteorological data. Over the years, the increasing availability of High-Performance Computing (HPC) resources has facilitated significantly the use of ensemble approaches across scientific and engineering domains (Wu and Levinson, 2021), particularly in weather forecasting (Toth and Kalnay, 1997; Leutbecher and Palmer, 2008; Bauer et al., 2015). Volcanic cloud forecasting has also benefited from ensemble modelling approaches, especially in areas like uncertainty quantification, hazard assessment and data assimilation. With the help of ensembles, it has been possible to characterise and quantify forecast uncertainties due to poorly constrained input parameters, model errors or uncertain weather data (Stefanescu et al., 2014), expanding upon and beyond traditional forecast products based on deterministic approaches (Folch et al., 2022). This is also true for volcanic hazard assessment, another aspect of paramount importance to support long-term risk-based decision-making and short-term emergency management during volcanic crises (Sandri et al., 2025). Probabilistic hazard assessments, in particular, require estimating the probabilities that a given airspace/location will be impacted by a volcanic cloud/fallout (Marzocchi et al., 2006), and a proper probabilistic hazard assessment usually require a large number of numerical simulations (Madankan et al., 2014; Titos et al., 2022). On the other hand, forecast errors can be reduced by incorporating observations into numerical models using ensemble-based data assimilation (Kalnay, 2003). Data Assimilation (DA) techniques have been widely used to study and forecast geophysical systems and have been applied in a variety of research and operational settings (Carrassi et al., 2018). In particular, several ensemble-based methods have been applied to volcanic plume forecasting (Folch and Mingari, 2023), including ensemble Kalman filter methods (Fu et al., 2016, 2017; Osores et al., 2020; Pardini et al., 2020; Mingari et al., 2022) and ensemble particle filtering (Zidikheri and Lucas, 2021a, b).

In all these application examples, the use of large ensembles can be computationally costly, even on HPC clusters, especially in the context of real-time services. Operational forecast frameworks must accommodate ensemble sizes to the available computing resources, resulting in ensembles limited to a few tens of members and the corresponding introduction of significant sampling errors (Weisheimer et al., 2014; Berner et al., 2017; Govett et al., 2024). In contrast, generative AI models have the potential to generate thousands of samples almost instantaneously at very low computational cost. In particular, Machine Learning (ML) algorithms can be trained to generate results consistent with the underlying physical laws, such as those describing atmospheric dispersion and transport of volcanic plumes. Some examples of physics-informed generative models preserving the underlying governing physical laws include Physics-Informed Neural Networks (PINNs) (Raissi et al., 2019; Kashinath et al., 2021), Diffusion Models with physical constraints (Gao et al., 2023; Wang et al., 2025), and Generative Adversarial Network (GANs) augmented with conservation laws or physical constraints (Yang et al., 2020; Tretiak et al., 2022). Similarly, the Variarional Autoencoder or VAE (Kingma and Welling, 2013) represents one of the most prominent examples of generative models. A VAE is a type of neural network that can learn a probabilistic mapping from a high-dimensional data space to a continuous latent space, typically of lower dimensionality than the original data space, enabling the reconstruction and synthesis of new states by sampling from this learned representation. Several recent studies have demonstrated the benefits of using a

https://doi.org/10.5194/egusphere-2025-6097
Preprint. Discussion started: 30 December 2025

low-dimensional latent space for data assimilation purposes (Pasmans et al., 2025). For example, Chen et al. (2025) trained a VAE using sea ice reanalysis data and conducted assimilation experiments with a particle-filter technique, taking advantage of a VAE to extract low-dimensional representations of sea ice physical fields. In this way, the authors significantly reduced model errors by assimilating satellite observations of sea ice concentration and thickness, demonstrating that the proposed methodology is a promising solution for nonlinear data assimilation in realistic high-dimensional geoscience applications. Alternatively, Pasmans et al. (2025) proposed a variant of the ensemble transform Kalman filter (Bishop et al., 2001) in which the analysis update is applied in the latent space of a VAE. In addition, they also showed that by introducing a second latent space for the observational innovations, the performance can be further improved when non-Gaussian observation errors are present.

Previous studies employing AI-based approaches to volcanic ash dispersion have focused primarily on the detection and segmentation of volcanic clouds in satellite imagery (Torrisi et al., 2024; Corradino et al., 2024). However, the use of generative AI techniques for volcanic-cloud modelling and forecasting remains largely unexplored. This work introduces a generative AI framework for statistically augmenting a small ensemble of physics-based numerical simulations, thereby enriching the ensemble forecast representation. A conventional VAE is employed to generate thousands of predictions (samples) of volcanic ash concentration. Specifically, an ensemble forecast of two-dimensional mass loading, i.e. vertical column mass per unit area of airborne ash, served as the training data for a convolutional neural network. The training, validation and test datasets were simulated using the VATD model FALL3D (Folch et al., 2020; Prata et al., 2021; Folch et al., 2025). By perturbing eruption source parameters and meteorological fields, FALL3D can produce an ensemble of simulations in a very efficient way by running a parallel job with multiple GPUs (Folch et al., 2022). As a first step, this paper focuses on characterising the statistical properties of VAE-generated ensembles and evaluating the VAE performance, while the study on using the latent space for satellite DA via deep learning is postponed to a following paper. The manuscript is organised as follows. Section 2 describes the FALL3D numerical model, the model configuration and the datasets used for training and performance evaluation. The neural network architecture is described in Sect. 3 along with a brief overview of the theoretical background behind VAEs. The results are presented in Sect. 4 and the performance is evaluated in terms of different metrics. Finally, conclusions are drawn in Sect. 5, where some implications of this work and possible future applications are discussed.

2 Physical model and datasets

A VATD model was used to produce the datasets required for VAE training and evaluation. Specifically, each dataset consists of an ensemble of numerical simulations performed using the latest release (version 9.1) of the open-source code FALL3D (Folch et al., 2025). FALL3D is an Eulerian model for atmospheric passive transport and deposition of different atmospheric species (e.g., volcanic ash and gases, mineral dust and radionuclides) based on the so-called Advection-Diffusion-Sedimentation (ADS) equation (Folch et al., 2020; Prata et al., 2021). The numerical model is GPU-accelerated using the OpenACC parallel programming model and has been designed to support increasingly larger scientific workloads in preparation for the transition to extreme-scale computing systems (Folch et al., 2023). The model has been deployed and optimised on the main pre-exascale

https://doi.org/10.5194/egusphere-2025-6097
Preprint. Discussion started: 30 December 2025

supercomputer systems in Europe: LUMI (Finland), Leonardo (Italy), and MareNostrum-5 (Spain), covering both NVIDIA and AMD GPU architectures.

FALL3D is designed to run ensemble simulations as a single, massively parallel job, allowing to maximize computational efficiency and scalability (Folch et al., 2022). Broadly speaking, ensemble modelling allows to characterise and quantify model uncertainties due to poorly constrained input parameters and errors in the numerical model, physical parameterisations or underlying model-driving meteorological data. In addition, the ability to generate ensemble runs makes it possible to improve forecasts by incorporating observations using different ensemble-based data assimilation techniques. In this work, different ensembles were simulated for training, validating and testing neural networks. In particular, we used model data from the recent EU-MODEX Tenerife 2025, organised under the European Civil Protection Mechanism. This exercise represented the first large-scale eruption drill in Spain to practice the response to a potential volcanic eruption in Garachico volcano (Tenerife, Canary Islands) and test evacuation, social assistance and communication during a volcanic crisis. On 26 September 2025, we provided real-time ash-dispersion forecasts for a hypothetical eruption scenario to support emergency decision-making and, afterwards, the results were used to explore the potential of VAE for ensemble expansion. As this study aims at generating ensembles expanding upon numerical simulations and to generate good approximations of their distribution, the distinction between real and hypothetical eruption scenarios is not relevant here.

2.1 Ensemble of numerical simulations

The configuration of the FALL3D model is summarised in Table 1. A 24-h forecast was performed assuming a continuous emission starting at 09:00UTC. A local-scale computational domain with a horizontal resolution of 0.1° and $120 \times 100 \times 25$ grid cells was defined, and the FALL3D dispersal model was driven by the Global Forecast System (GFS) weather forecast from the National Centers for Environmental Prediction (NCEP).

Table 1. FALL3D model configuration.

Parameter	Value
Start time	26 September 09:00UTC
Run time	24 h
Horizontal grid resolution	$0.1^\circ \times 0.1^\circ$
Number of grid points	$120 \times 100 \times 25$
Grain Size Distribution	bi-Gaussian
Emission source	Suzuki source (Pfeiffer et al., 2005)
Mass emission rate	Estimated (Woodhouse et al., 2016)
Ensemble size	64, 256 and 2048

Each ensemble was simulated by perturbing the Eruption Source Parameters (ESPs) and wind components using a Latin Hypercube Sampling (LHS, McKay et al., 1979) to efficiently sample the parameter space. Table 2 lists the reference value

https://doi.org/10.5194/egusphere-2025-6097
Preprint. Discussion started: 30 December 2025

and sampling range for each perturbed parameter. The top height of the eruption column was sampled around the central value $H = 9$ km avl (above the vent level) with a perturbation range of 30% and was assumed to be constant during the simulations for each ensemble member. The vertical distribution of the source term was also modified by perturbing the Suzuki parameter A_s controlling the maximum emission height (see Folch et al., 2020, for more information). In addition, the horizontal wind components were perturbed considering a sampling range of 25%. Eruptive column height and wind are widely reconsigned as one of the most sensitive parameters/inputs for volcanic ash transport models (Devenish et al., 2012; Poulidis et al., 2018).

Table 2. Ensemble configuration. The perturbed model parameters are: eruption column height (H), mass eruptive rate (MER), Suzuki parameter A_s and the wind horizontal components.

Parameter	Reference value	Distribution	Sampling range
H	9 km avl[†]	Uniform	±30%
MER	$\sim 1.6 \times 10^6$ kg/s[‡]	Uniform	±25%
A_s	4	Uniform	1
Winds	GFS[§]	Uniform	±25%

[†] Height given in km above the vent level;
[‡] Estimated from H according to Woodhouse et al. (2016);
[§] Global Forecast System (GFS) weather forecast from NCEP.

Three ensemble forecasts were simulated to generate the training (simulation A), validation (simulation B), and test (simulation C) datasets. As summarized in Table 3, the ensemble corresponding to simulation A, consisting of 256 members, was used to train the VAE. The 64-member ensemble (simulation B) was used during the training step to monitor the performance and tuning of the neural network hyperparameters. Finally, the larger 2048-member ensemble (simulation C) was used only at the end as a purely physics-based reference to evaluate the VAE model, providing a fair (unbiased) report of model accuracy. All numerical simulations were performed on the LUMI EuroHPC supercomputing system, hosted at CSC's (Finland). In particular, simulation C required substantial HPC resources and was executed on the LUMI-G partition allocating 256 computing nodes and a total of 2048 GPUs.

Table 3. Datasets built from ensemble of numerical simulations.

Simulation	Ensemble size	Purpose
A	256	Training
B	64	Validation
C	2048	Test

https://doi.org/10.5194/egusphere-2025-6097
Preprint. Discussion started: 30 December 2025

3 Variational Autoencoder

A Variational Autoencoder (VAE) is a type of generative ML model based on a neural network architecture capable of learning an efficient and probabilistic encoding of data into a latent space of low dimensionality (Kingma and Welling, 2013). By sampling vectors $\boldsymbol{z}$ from the learned distribution in the latent space, a decoder can afterwards generate new and realistic data samples, even when the network was trained to encode complex and high-dimensional input data. The scheme in Fig. 1 shows a typical architecture of a VAE composed of convolutional layers. First, the encoder maps input vectors $\boldsymbol{x}$ to the parameters of a multi-dimensional probability distribution over the latent space, e.g. a Gaussian with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ (note that bold symbols denote vector quantities). Second, a vector $\boldsymbol{z}$ is sampled in the latent space and passed to the decoder, which aims to reconstruct the original input $\boldsymbol{x}$.

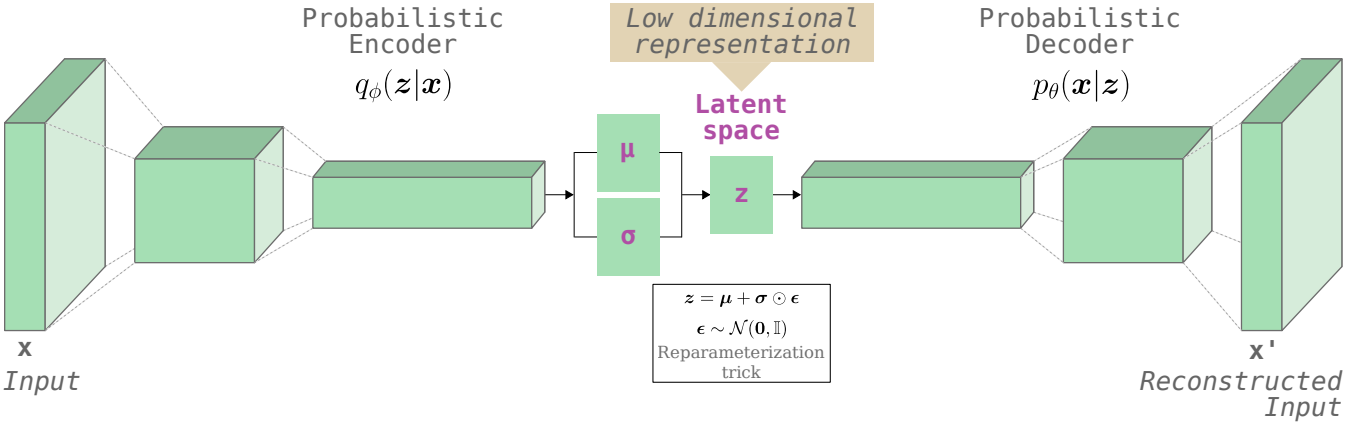

Figure 1. The basic scheme of a VAE composed of two-dimensional convolutional layers. The encoder compresses an input $\boldsymbol{x}$ into the low-dimensional latent space. The decoder attempts to reconstruct the original data from a vector $\boldsymbol{z}$ sampled from the latent space.

The training of the VAE is guided by a loss function that balances the fidelity of the data reconstruction against a term that regularizes the probability distributions in the latent space towards a simple prior distribution $p(\boldsymbol{z})$:

$$\mathcal{L}_{VAE} = -\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\right] + D_{KL}\left(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \| p(\boldsymbol{z})\right) \tag{1}$$

where $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ is the encoder approximate posterior and $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ the decoder likelihood. Equivalently, it can be expressed as the sum of two terms:

$$\mathcal{L}_{VAE} = \mathcal{L}_R + \mathcal{L}_{KL}$$

The first term ($\mathcal{L}_R$, reconstruction loss) encourages reconstructions to be accurate and, in this work, it is assumed to be the Mean Squared Error (MSE) between the input ($\boldsymbol{x}$) and its reconstructed output ($\boldsymbol{x}$'):

$$\mathcal{L}_R = \frac{1}{N}\sum_{i=1}^{N}(x_i - x_i')^2 \tag{2}$$

https://doi.org/10.5194/egusphere-2025-6097
Preprint. Discussion started: 30 December 2025

being N the dimension of the original space. The second term ($\mathcal{L}_{KL}$) is defined in terms of the Kullback-Leibler (KL) divergence, D_{KL}, and measures the difference between the approximate posterior $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and the prior distribution $p(\boldsymbol{z})$. In practice, the D_{KL} term forces q_ϕ to be close to the prior distribution $p(\boldsymbol{z})$. Traditionally, the standard VAE assumes a diagonal multivariate normal distribution for the approximate posterior, $q_\phi(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma^2}\mathbb{I})$, and a standard normal distribution for the prior, $p(\boldsymbol{z}) = \mathcal{N}(\mathbf{0}, \mathbb{I})$, leading to:

$$\mathcal{L}_{KL} = -\frac{1}{2}\sum_{i=1}^{K}\left[1 + \log\sigma_i^2 - \mu_i^2 - \sigma_i^2\right] \tag{3}$$

being $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)^\intercal$ the vector of means, $\boldsymbol{\sigma^2} = (\sigma_1^2, \ldots, \sigma_K^2)$ the vector of variances and K the dimension of the latent space where vectors $\boldsymbol{z}$ live. Finally, the so-called β-VAE is a modification of the standard VAE that uses a hyperparameter β to weight the KL divergence term in the loss function:

$$\mathcal{L}_{VAE} = \mathcal{L}_R + \beta\mathcal{L}_{KL} \tag{4}$$

This is the expression of the loss function used in this work. In summary, VAE training optimises the encoder and decoder by minimising the VAE loss function (4) on a training set of input vectors $\{\boldsymbol{x_i}\}$, so that inputs are accurately reconstructed from latent samples, while the latent distribution remains close to the prior.

3.1 Training

A VAE was implemented in the PyTorch framework (Paszke et al., 2019) using an architecture based on two-dimensional convolutional layers with kernel size of 3×3. The VAE configuration is defined by the hyperparameters reported in Table 4. The resulting number of trainable parameters was 698,849. The neural network was trained in batches using the Adam optimizer for 400 epochs. The reconstruction loss and KL divergence (averaged over batches) were monitored at each epoch for both the training and validation datasets to prevent overfitting. The training (validation) dataset is an ensemble composed of 256 (64) two-dimensional fields of volcanic ash column mass on a grid of size 120×100. Note also that, for the purpose of this work and without loss of generality, training was done on a single forecast time (+24h). The inclusion of time dependency is straightforward and can be done by simply expanding the dimension of the training dataset by adding successive timestamps. As outlined in the scheme of Fig. 1, the encoder network maps the input data $\boldsymbol{x} \in \mathbb{R}^{120\times100}$ through three successive convolutional layers, followed by a flattening operation to two distinct latent space parameter vectors $\boldsymbol{\mu}$ (mean) and $\boldsymbol{\sigma}^2$ (variance or log-variance) of dimension $K = 16$. In other words, each component z_i of the latent vector ($\boldsymbol{z}$) is modelled as independently following a Gaussian distribution with parameters μ_i and σ_i^2, while the KL term encourages these distributions to be close to the standard normal prior $\mathcal{N}(0,1)$. The decoder then reconstructs the input $\boldsymbol{x'}$ from the sampled latent vector $\boldsymbol{z} \in \mathbb{R}^{16}$ using a fully connected layer followed by three transposed convolutional layers (or deconvolutional layers), restoring the original dimensions, i.e. $\boldsymbol{x'} \in \mathbb{R}^{120\times100}$.

By construction, the components of the latent vector encoded from the training dataset should be closely distributed according to independent standard normal distributions: $z_k \sim \mathbb{N}(0,1)$. The covariance matrix (Fig. 2a) indicates a very low correlation

https://doi.org/10.5194/egusphere-2025-6097
Preprint. Discussion started: 30 December 2025

Table 4. Hyperparameters used for the VAE training.

Hyperparameter	Value
Latent dimension (K)	16
$\beta^{\dagger}$	6
Data size (N)	120×100
Number of epochs	400
Batch size	16
Convolutional layers	6
Output channels	16, 32 and 64
Kernel Size	3×3
Stride	1×1
Activation function	ReLU
Learning rate	1E-3
Optimizer	Adam
Normalization	Percentile-based scaling

† Scaling factor for the KL divergence term in the β-VAE formulation, defined in Eq. (4).

between each pair (z_i, z_j), as expected. It is also interesting to corroborate whether the test dataset is encoded in the same region of the latent space covered by the training dataset. Figure 2b shows a two-dimensional visualization of the latent space based on the Uniform Manifold Approximation and Projection (UMAP) algorithm (McInnes et al., 2020). This is a technique for non-linear dimensionality reduction specifically designed to preserve the local structure of the data. As observed, the test and training datasets cover the same region of the UMAP projection, exactly as it occurs in the original latent space (this can be verified by plotting any pair of components of the latent vector, e.g. z_5 vs z_4). These remarks indicate, as a whole, that the network was trained correctly and works as expected. In addition, it is also worth showing how the encoding can discriminate between qualitatively different model solutions (ensemble members) by grouping them in different regions of the latent space. This is illustrated in Fig. 2b by means of coloured-scale dots, where members with eruption column heights higher and lower than the reference (ensemble central member) are clearly concentrated in disjoint (i.e. non-overlapping) regions of latent space. In other words, the VAE encoding implicitly produces a latent space with clustering-like structure.

3.2 Evaluation metrics

As long as the neural network was trained using a physical model (simulation A, 256 members), a key question is whether an ensemble of new samples $\{\boldsymbol{x_i}\}$ generated by the VAE resembles the distribution of a larger ensemble produced by the same physical model $\{\hat{\boldsymbol{x}}_{\boldsymbol{i}}\}$ (simulation C, 2048 members). In other words, could a VAE trained with a small ensemble be used to generate a much larger ensemble indistinguishable from its physics-based equivalent? In order to answer this question,

https://doi.org/10.5194/egusphere-2025-6097
Preprint. Discussion started: 30 December 2025

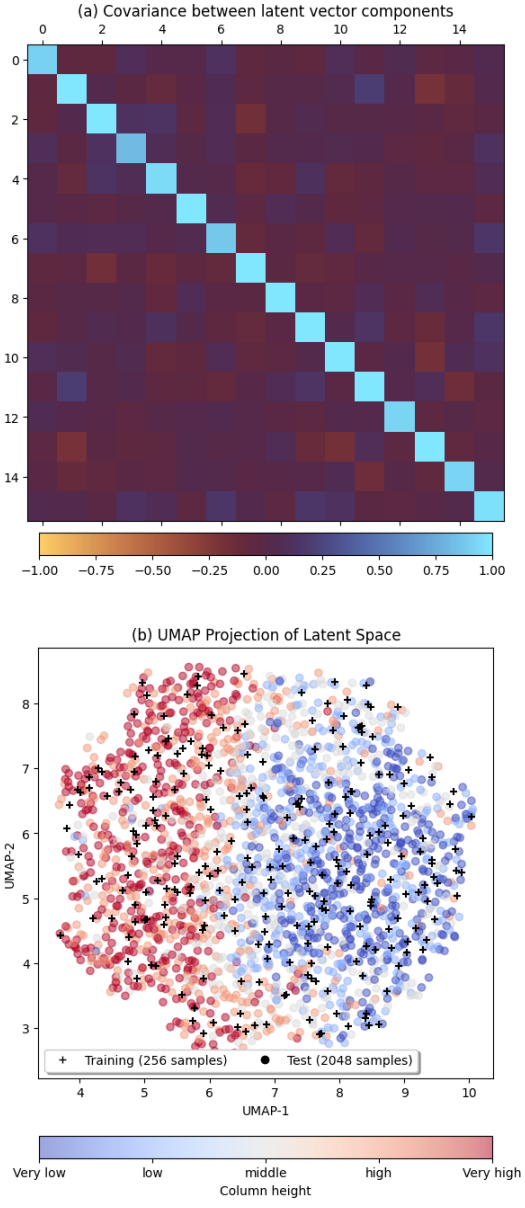

Figure 2. Encoding of the training and test datasets in the latent space. The independence between the distributions of the latent vector components can be confirmed from the covariance matrix (a). The test dataset is encoded in the same region of the latent space covered by the training dataset, as shown in the UMAP two-dimensional visualization of the latent space (a).

probabilistic distributions or sample statistics, such as ensemble mean, were compared against the reference ensemble (test dataset) consisting of 2048 simulations. Specifically, we first train a generative AI model using a 256-member ensemble, then

https://doi.org/10.5194/egusphere-2025-6097
Preprint. Discussion started: 30 December 2025

use the VAE to generate a new ensemble of synthetic samples, and finally evaluate their performance against the 2048-member physics-based ensemble.

Several performance metrics can be considered to quantify the ability of the generative AI model to produce realistic ash concentration outputs. For example, the Root-Mean-Square Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{y}_j)^2} \tag{5}$$

can be evaluated in terms of the ensemble means, i.e. $\boldsymbol{y} = 1/M \sum \boldsymbol{x_i}$ and $\hat{\boldsymbol{y}} = 1/\hat{M} \sum \hat{\boldsymbol{x}}_{\boldsymbol{i}}$ where, in our particular case, $\hat{M} = 2048$ and M is the number of generated samples. Similarly, the Mean Bias Error (MBE) is computed as:

$$\text{MBE} = \frac{1}{N}\sum_{j=1}^{N} y_j - \hat{y}_j \tag{6}$$

The Mean Absolute Percentage Error (MAPE) provides another relative measure of prediction accuracy:

$$\text{MAPE} = 100 \times \sum_{j=1}^{N} \left| \frac{y_j - \hat{y}_j}{\hat{y}_j} \right| \tag{7}$$

In addition to the standard performance metrics introduced above, it is important to evaluate how accurately the VAE reproduces the underlying target distribution. The Jensen-Shannon (JS) divergence is a measure of the similarity between two probability distributions. Specifically, for two distributions $P = \{p_i\}$ and $Q = \{q_i\}$, the JS divergence is a symmetrized version of the KL divergence:

$$JSD(P\|Q) = \frac{1}{2}D_{KL}(P\|M) + \frac{1}{2}D_{KL}(Q\|M) \tag{8}$$

where $M = \frac{1}{2}(P+Q)$ is a mixture distribution of P and Q, while D_{KL} can be interpreted as the expected logarithmic difference between the probability distributions (Goodfellow et al., 2016):

$$D_{KL}(P\|Q) = \sum p_i \log \frac{p_i}{q_i} \tag{9}$$

The discrete distributions in (9) are estimated from histograms in this work.

Finally, the two-dimensional isotropic power spectrum, $P(k)$, is computed to assess the ability of the VAE to generate realistic samples over different spatial scales, in consistency with physically-based models. It is defined as the radially-averaged 2D Power Spectrum Density (PSD) and is a function of the wavenumber magnitude, k:

$$P(k) = \frac{1}{N_k}\sum PSD(k_x, k_y) \tag{10}$$

where $k = \sqrt{k_x^2 + k_y^2}$ is the magnitude of the wavevector, $PSD(k_x, k_y)$ is the two-dimensional Power Spectral Density obtained from a Fast Fourier Transform (FFT), and N_k is the number of data points within an annular ring around k in the Fourier space.

https://doi.org/10.5194/egusphere-2025-6097
Preprint. Discussion started: 30 December 2025

4 Results and discussion

Once the VAE neural network has been trained with numerical simulations, it is possible to generate at very low computational cost new physics-grounded samples, which are expected to be consistent with the underlying physical laws that govern atmospheric dispersion. For illustrative purposes, Fig. 3 shows a VAE-generated ensemble consisting of $M = 12$ samples. The samples represent two-dimensional ash column mass plots, i.e. the vertically integrated ash mass per unit area. The variability in the ensemble is a consequence of the different eruptive source parameters (e.g., eruptive column height) and meteorological fields used to simulate the training dataset. For example, samples in Fig. 3c,d correspond to cases with high eruptive column heights and strong emission, whereas other samples are associated with low column heights and weak emission (e.g. Fig. 3h). The volcanic plume is advected eastwards from the vent (16.76°W, 28.27°N) in most cases. Additionally, some cases are characterised by a secondary plume with a SE- or even W-component in the transport pattern, suggesting the presence of strong wind shear that introduces variability into the ensemble and leads to qualitatively different solutions.

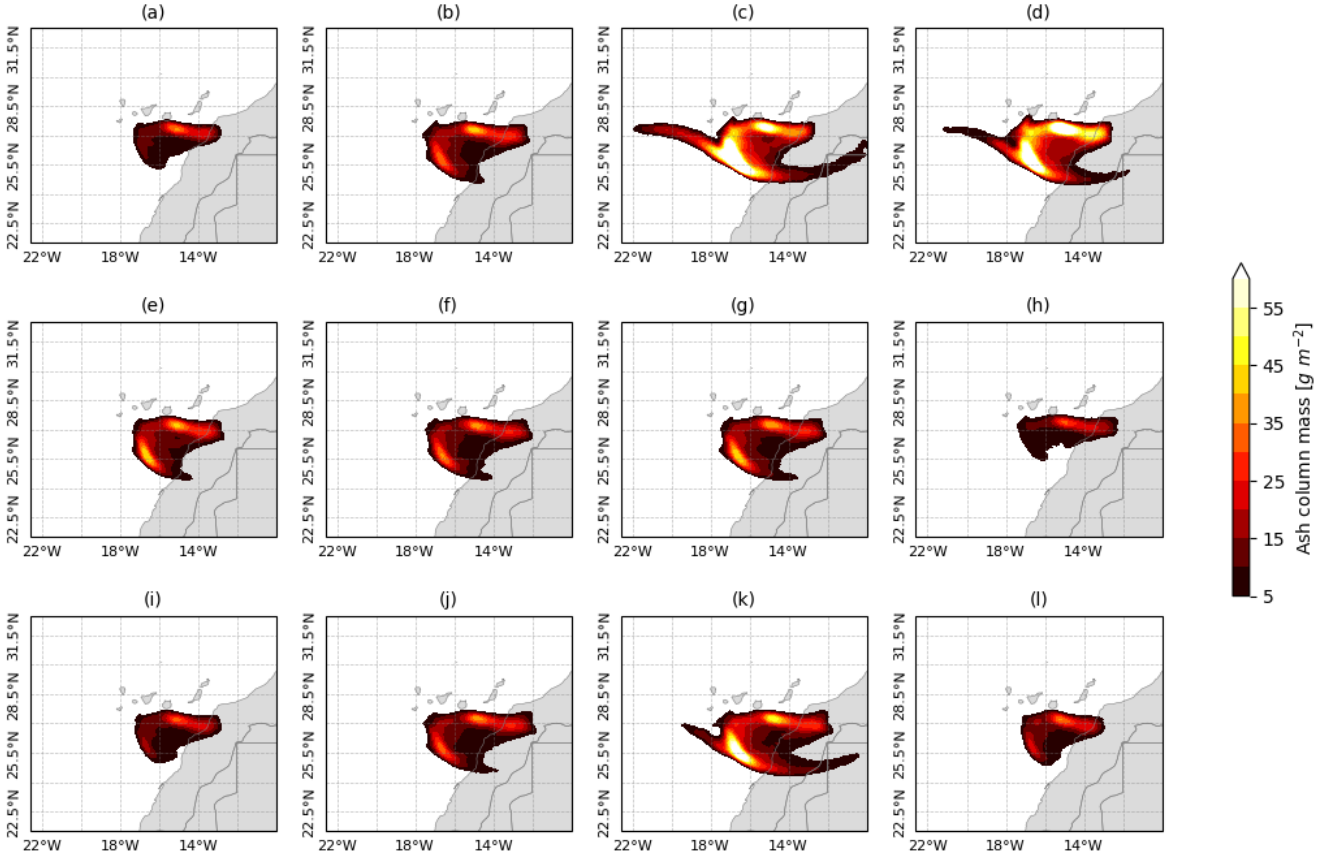

Figure 3. The neural network was trained to generate samples of two-dimensional ash column mass, i.e. the total mass of volcanic ash in a vertical column. In this example, a VAE-generated ensemble composed of 12 samples is shown.

https://doi.org/10.5194/egusphere-2025-6097
Preprint. Discussion started: 30 December 2025

4.1 Sensitivity study

A sensitivity test was conducted to evaluate the performance of different neural network configurations. Performance is evaluated in terms of the RMSE of the ensemble mean using (5) (see Sect. 3.2). In other words, the RMSE reported in this section measures how far the mean of the VAE-generated ensembles deviates from the ensemble mean of the test dataset. We trained multiple VAEs changing two key hyperparameters: the latent space dimension K and the factor β in (4), while fixing the remaining hyperparameters with the values specified in Table 4. Since the VAE is a probabilistic model, each VAE-generated ensemble leads to different performances. In order to account for this stochastic variability, several 2048-member ensembles were generated. For each combination of hyperparameters tested, 20 ensembles were generated and the boxplot in Fig. 4 integrates the resulting values of RMSE. In all cases, the network was trained for 400 epochs, except for configurations with latent dimensions of $K = 4$ and $K = 2$, where early stopping was required to avoid overfitting. According to the trend suggested by Fig. 4, the best performance is achieved for latent dimensions $K \gtrsim 6$ and $\beta \gtrsim 6$. Remarkably, performance degrades slightly for large latent space dimensions (Fig. 4a). There are two possible explanations for this anomaly. First, large latent spaces often produce less stable training dynamics and worse generalization: enlarging latent spaces introduces more degrees of freedom than the data can realistically support. When the dimensionality grows beyond what is needed to represent the underlying factors, the model may spread information across latent variables in an inefficient way. This leads to poor compression, weaker regularization and a latent representation that becomes harder to structure during training, limiting the extraction of compact and meaningful representations. As a result, the decoder receives a less informative latent code, reducing the reconstruction quality. On the other hand, increasing the latent dimensionality can slow convergence: since the model must explore a much larger parameter space before reaching a stable configuration, it requires more training iterations and makes optimisation more difficult. In addition, Fig. 4 shows the impact of the β parameter, which controls the strength of the KL regularization in the loss function and, therefore, the balance between reconstruction fidelity and latent-space structure. Higher values force the model to impose a tighter, more organised latent space, at the expense of losing some fine-scale details in the reconstructions. In our case, $\beta \approx 6$ suggests that a strongly regularised latent representation is more beneficial for the task than capturing every small feature in the input data. This indicates that the model gains more from enforcing a well-structured latent space than from preserving high-resolution details during reconstruction. This is clear in Fig. 2a, where the latent-space covariance matrix is nearly diagonal, indicating that the latent variables are uncorrelated. This structure suggests that each dimension captures a distinct component of the underlying variability, which is consistent with a more disentangled and well-regularized latent representation. Further details are provided in Sect. 4.4.

The results presented below are based on the VAE trained with the hyperparameters specified in Table 4, i.e. using a latent dimension of $K = 16$ and $\beta = 6$.

4.2 Ensemble mean and spread

The VAE ability to reproduce sample statistics, such as the ensemble mean and spread (standard deviation), is examined in this section, where a few metrics are reported. For example, Fig. 5 shows the ensemble mean for the test dataset (target). Ideally,

https://doi.org/10.5194/egusphere-2025-6097
Preprint. Discussion started: 30 December 2025

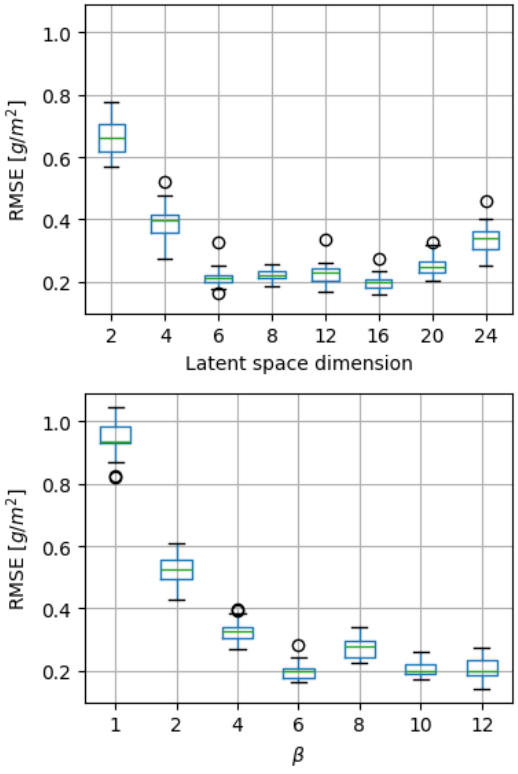

Figure 4. VAE sensitivity analysis for different combinations of two hyperparameters: the latent space dimension K (top) and the β parameter (bottom), defined in (1). The performance is measured in terms of the Root-Mean-Square Error (RMSE), computed from the ensemble mean.

VAE should closely approximate this target to achieve good performances. Similarly, Fig. 6 shows the corresponding ensemble mean for the training dataset (Fig. 6a) and VAE (Fig. 6b). In both cases, the target is correctly approximated. Deviations from the target are quantified by bias maps. In relative terms, the bias is small for both the training dataset (Fig. 6c) and VAE (Fig. 6d), although the training dataset provides a slightly better estimate of the ensemble mean in general.

A summary of the metrics is shown in Table 5. Metrics are computed for multiple VAE-generated ensembles composed of $M = 5000$ samples in order to account for the variability introduced by stochastic sampling. The average over 20 ensembles is reported along with the standard deviation for the uncertainty. The target test dataset is taken as a reference in order to evaluate each of the metrics, as mentioned in Sect. 3.2. The VAE errors measured in terms of the RMSE are twice as high as those evaluated from the training dataset. In relative terms, this represents an error of 1.78% (training) and 3.4% (VAE) according to the MAPE evaluated from the ensemble mean. In order to avoid divisions by zero and numerical instabilities, the MAPE was actually computed considering only the region where the test dataset is above $5\ gm^{-2}$ (i.e. the region shown in Fig. 5). We can conclude that VAE is properly capturing the first and second moments of the distribution, although the training estimations for

https://doi.org/10.5194/egusphere-2025-6097
Preprint. Discussion started: 30 December 2025

the ensemble mean and spread are still slightly better. However, next section shows how the VAE achieves a better performance for other statistics more sensitive to sampling errors when large enough samples are generated.

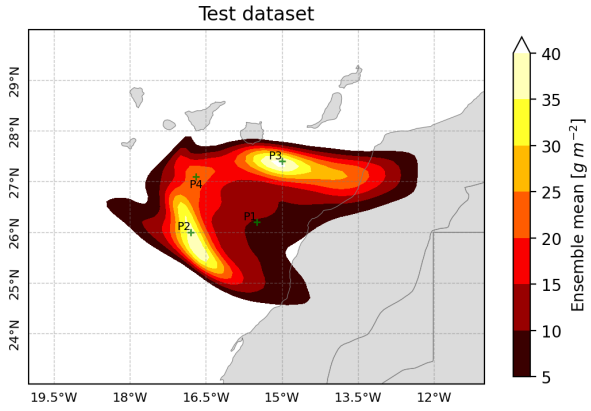

Figure 5. Ensemble mean computed from the test dataset, consisting of 2048 physics-based numerical simulations generated by the FALL3D dispersal model.

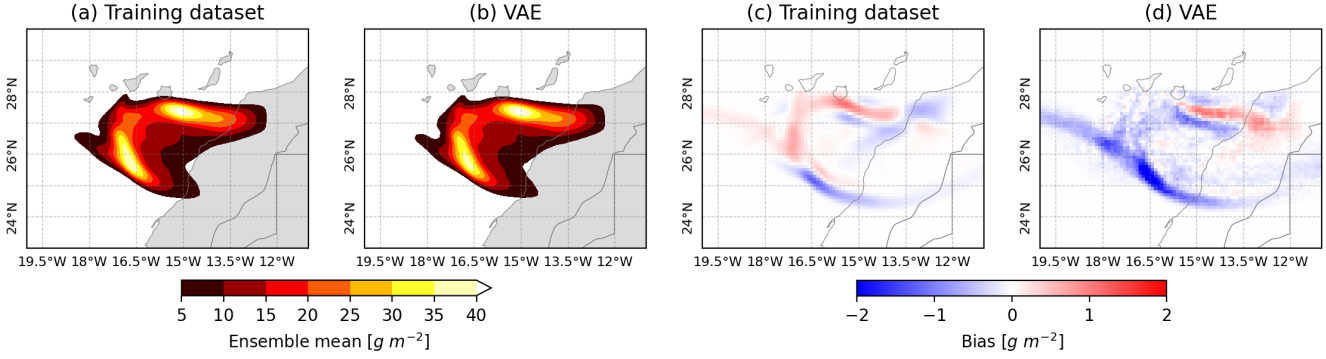

Figure 6. Ensemble mean maps obtained from the training dataset (a) and a VAE-generated ensemble with 5000 samples (b). The bias maps for the training (c) and VAE (d) ensemble means are computed by subtracting the target (test dataset) results.

4.3 Probability distributions

280 An important aspect in hazard and impact assessment is to evaluate exceedance probabilities, i.e. the probability that a specific hazardous threshold is exceeded within a given time frame. In the context of ensemble modelling, probability distribution estimations can suffer from sampling errors, particularly when small ensembles are considered. Figure 7 presents the exceedance probabilities for three different thresholds, demonstrating that the VAE is capable of generating accurate probability distributions as well as realistic samples.

https://doi.org/10.5194/egusphere-2025-6097
Preprint. Discussion started: 30 December 2025

Table 5. A summary of some performance metrics evaluated using the test dataset as the reference target. The VAE metrics were averaged over 20 ensembles generated to account for stochastic variability. The uncertainty range is given by the standard deviation.

Metric	Variable	Training dataset	VAE (2048 samples)
RMSE	Mean	0.108	0.20 ± 0.03
MBE	Mean	0.001	-0.04 ± 0.02
MAPE (%)	Mean	1.784	3.4 ± 0.6
RMSE	Spread	0.255	0.56 ± 0.04
MBE	Spread	-0.004	-0.17 ± 0.01

In units of gm^{-2}, except MAPE.

In order to examine whether the sampled ensemble provides a good approximation to the actual distribution (target), a few histograms are shown in Fig. 8 at four locations (see Fig. 5), comparing results for the training dataset (left column) and a VAE ensemble with 5000 samples (right column), considering 128 bins between the minimum and maximum values for each location. As expected, the distribution estimated from the training dataset with a small ensemble size (256) results in a very noisy distribution, while the VAE distribution provides a good approximation to the test dataset distribution, even for multimodal distributions (e.g., Fig. 8a).

Naturally, the VAE capacity to approximate the distribution depends on the number of samples generated. In order to assess the quality of the approximation for different amounts of samples, the JS divergence was computed from discrete probability distributions using (8). Specifically, JSD was obtained from histograms with a fixed range (128 bins between 0 and 40 $g\ m^{-2}$) for each grid cell. As a measure of the worst-case conditions, the 99th percentile of the grid-cell values is reported in Fig. 9, taken as a robust estimation of the maximum JSD over the domain. The results for multiple VAE-generated ensemble sizes are compared with the training dataset. As expected, VAE improves the training distribution only for large ensemble sizes ($M \gtrsim 256$), reaching an asymptotic value of approximately $JSD \sim 0.042$, which represents a 2.6x improvement over the corresponding training metric ($JSD = 0.11$).

4.4 Isotropic power spectrum

The isotropic power spectrum, i.e. the radially averaged two-dimensional Power Spectral Density (PSD), was computed using (10). Figure 10 presents the results for the test dataset and VAE, averaged over multiple ensemble realizations (each ensemble with $M = 5000$ samples). It is worth noting that, instead of the classical energy cascade spectrum according to $\propto k^{-5/3}$, the test dataset exhibits an exponential decay (solid red line), likely due to the diffusion processes introduced by the dispersal model FALL3D. The generative model, with a strongly regularized latent space, reproduces well the large scales but fails at smaller scales, as shown in Fig. 10 (green triangles). The isotropic power spectrum deviates from the exponential decay for wavenumbers above $\sim 0.3\ \Delta x^{-1}$, approximately, where Δx is the grid size. In other words, the VAE is unable to reproduce variability at spatial scales below $\sim 3\ \Delta x$. The flat high-wavenumber tail in the 2D power spectrum seems to be related to the

https://doi.org/10.5194/egusphere-2025-6097
Preprint. Discussion started: 30 December 2025

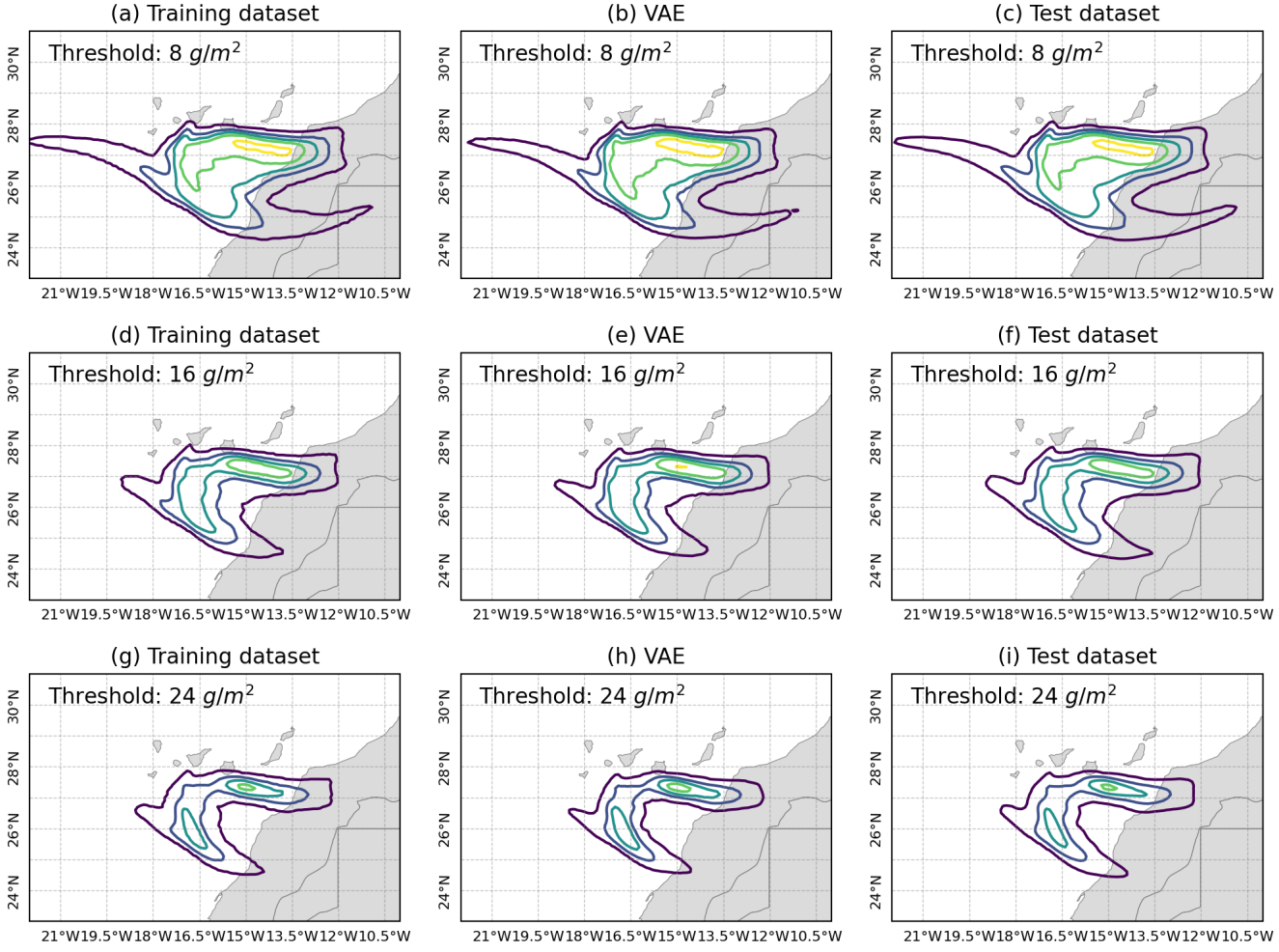

Figure 7. Exceedance probabilities for three ash column mass thresholds (8, 16 and 20 $g\ m^{-2}$ from top to bottom). The probability contours correspond to 2%, 25%, 50%, 75% and 98%. Results for the training dataset (left column) and a VAE-generated ensemble with 5000 samples (central column) are compared to the test dataset target (right column).

convolutional kernel size used in this work (i.e., 3×3) and the limitations of the decoder upsampling. Our decoder architecture is based on transposed convolution layers, and it is widely known that transposed convolutions usually lead to chequerboard artifacts. Actually, this effect can also be observed from the bias map of the ensemble mean (Fig. 6d), where a noisy pattern with relatively low intensity and high-frequency structure can be identified. A new neural network was trained using a larger kernel size (5×5) in order to investigate whether it yields to a more effective low-pass filtering (blue crosses in Fig. 10). Although the background noise was reduced slightly, the results remained almost identical. Future work should investigate more advanced neural network architectures to better address this limitation.

https://doi.org/10.5194/egusphere-2025-6097
Preprint. Discussion started: 30 December 2025

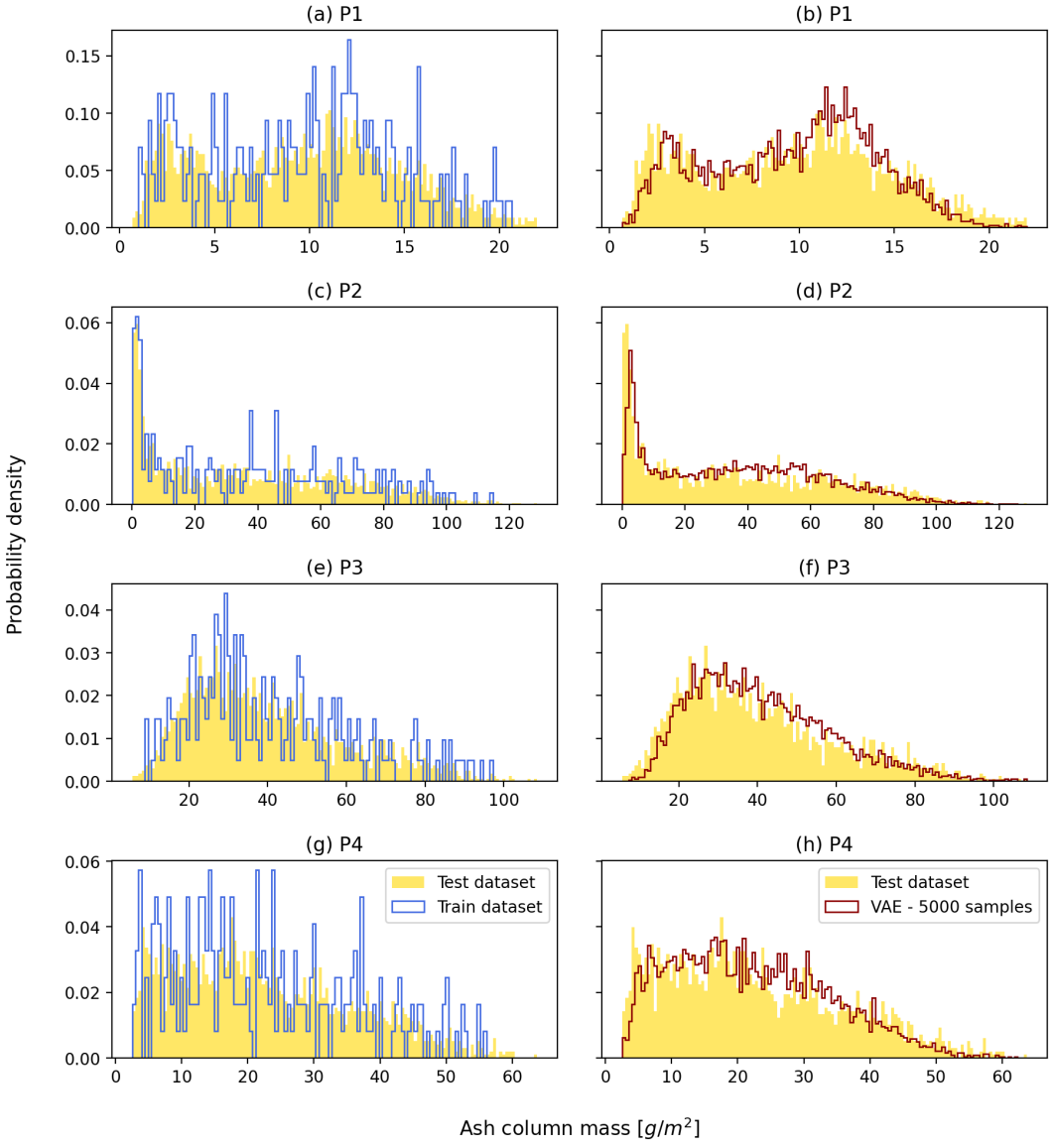

Figure 8. Histograms for the training datasets (256 samples, left column) and a VAE-generated ensemble (5000 samples, right column) at different locations. The results are compared against the test dataset (2048 samples), assumed to be the reference target (yellow shaded area). For illustrative purposes, four sampling locations (see Fig. 5) were selected: P1 (a,b), P2 (c,d), P3 (e,f), and P4 (g,h).

315 Finally, this behaviour is consistent with the disentangled latent representation enforced by a high value of the β hyper-parameter (see Fig. 4). In fact, the VAE learns a smooth, structured encoding that preserves the dominant generative factors while filtering out small-scale variations that would otherwise help reduce the reconstruction error. The effects of this strongly regularized latent space are also visible in the spectral analysis. While the VAE successfully reproduces the low-frequency

https://doi.org/10.5194/egusphere-2025-6097
Preprint. Discussion started: 30 December 2025

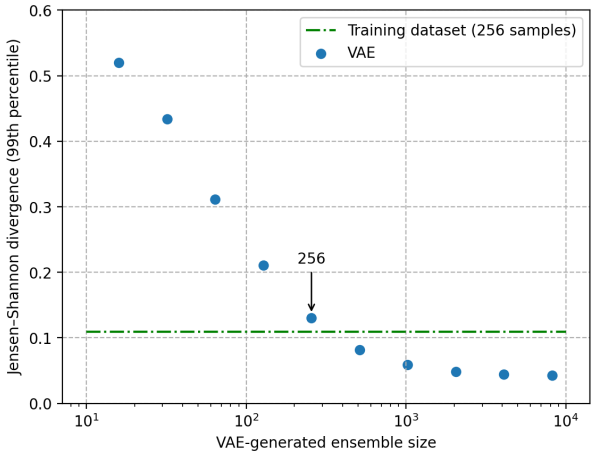

Figure 9. The Jensen-Shannon (JS) divergence was computed for each grid cell considering different sizes of the VAE-generated ensemble. The 99th percentile is reported as a worst-case measure of performance. Low values indicate a good approximation to the actual probability distribution. The result for the training dataset is also indicated for reference.

components associated with large-scale structures, it systematically underestimates the high-frequency content responsible for fine-scale details.

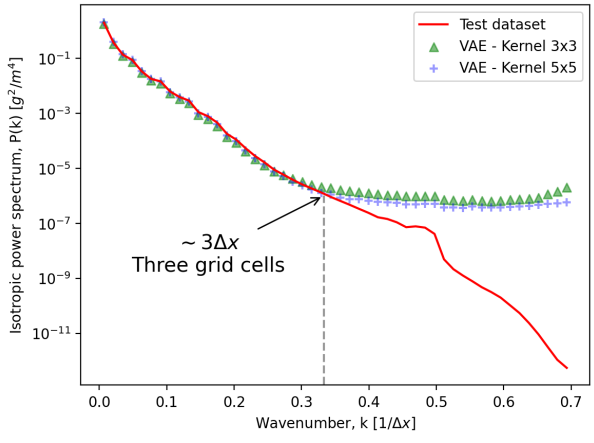

Figure 10. The isotropic power spectrum, i.e. radially averaged two-dimensional Power Spectral Density (PSD), reveals inconsistencies in the spatial structure of the VAE-generated samples. The generative model reproduces large scales but fails at scales smaller since the VAE is introducing artificial noise for large k. Since the wavenumber is expressed in terms of the grid size, Δx, we can conclude that the VAE is unable to reproduce the variability over spatial scales below $\sim 3\ \Delta x$, approximately.

https://doi.org/10.5194/egusphere-2025-6097
Preprint. Discussion started: 30 December 2025

5 Conclusions

This paper has untapped the use of generative AI modelling for atmospheric transport of volcanic clouds. As a pioneer step, we proposed a simple approach based on a convolutional Variational Autoencoder (VAE), a type of ML model capable of learning a probabilistic representation of data in a low dimensional latent space and generate new data (samples) similar to those simulated by a physical model.

Numerical simulations in geophysics can be highly demanding in terms of computational resources. For example, the ensemble simulation for the test dataset used in this work was run on the LUMI supercomputer using 2048 GPUs. With the proposed methodology, it is possible to generate thousands of new physics-grounded samples with statistical properties close to the true (model) in very short times using just a personal computer. Compared to the training dataset, VAE can provide a better approximation to the actual distribution with a 2.6x improvement in terms of the Jensen-Shannon (JS) divergence when large enough ensembles are generated. However, as expected, the training data exhibits superior performance to the VAE across nearly all statistical metrics considered in this work. However, the VAE-generated ensemble still provides a really good approximation of the ensemble mean, spread and probabilities of exceedance. For example, the typical VAE error in relative terms was found to be around 3.4% for the ensemble mean, while the training error was 1.78% on average. This is negligible compared to typical errors found when model forecasts are compared with observations (i.e., satellite retrievals or ground-based measurements) in real applications.

In summary, this methodology can be extremely useful for ensemble expansion, uncertainty quantification, and surrogate modelling, enabling to approximate the underlying distribution of the training dataset and generate synthetic predictions from it. Furthermore, by enabling latent space data assimilation via deep learning approaches, this methodology has the potential to dramatically reduce forecast errors in operational frameworks.

Code availability. FALL3D v9.1 is available under the version 3 of the GNU General Public License (GPL) at https://gitlab.com/fall3d-suite/fall3d/ and https://doi.org/10.5281/zenodo.17737081 (Folch et al., 2020; Prata et al., 2021).

Author contributions. Conceptualisation, L.M., M.T.; Methodology, L.M.; Software, L.M., A.F., H.P.; Resources, L.M.; Writing—original draft, L.M.; Writing—review and editing, L.M., A.F., H.P., M.T.; Visualisation, L.M.; Supervision, A.F.; Funding Acquisition, A.F. All authors have read and approved the final version of the manuscript.

Competing interests. The authors declare that they have no conflict of interest.

https://doi.org/10.5194/egusphere-2025-6097
Preprint. Discussion started: 30 December 2025

Acknowledgements. Funded by the European Union. This work has received funding from the European High Performance Computing Joint Undertaking (JU) and Spain, Italy, Iceland, Germany, Norway, France, Finland and Croatia under grant agreement No 101093038, ChEESE-2P, project PCI2022-134973-2 funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. Computational resources were provided by the LUMI supercomputer through the EuroHPC Joint Undertaking (project EHPC-DEV-2025D10-040) and MareNostrum 5 ACC supercomputer through the EuroHPC Joint Undertaking (project EHPC-REG-2024R02-128). LUMI is hosted by CSC - IT Center for Science and MareNostrum-5 is hosted by the Barcelona Supercomputing Center (BSC).

https://doi.org/10.5194/egusphere-2025-6097
Preprint. Discussion started: 30 December 2025

References

Bauer, P., Thorpe, A., and Brunet, G.: The quiet revolution of numerical weather prediction, Nature, 525, 47–55, 2015.

Berner, J., Achatz, U., Batté, L., Bengtsson, L., de la Cámara, A., Christensen, H. M., Colangeli, M., Coleman, D. R. B., Crommelin, D., Dolaptchiev, S. I., Franzke, C. L. E., Friederichs, P., Imkeller, P., Järvinen, H., Juricke, S., Kitsios, V., Lott, F., Lucarini, V., Mahajan, S., Palmer, T. N., Penland, C., Sakradzija, M., von Storch, J.-S., Weisheimer, A., Weniger, M., Williams, P. D., and Yano, J.-I.: Stochastic Parameterization: Toward a New View of Weather and Climate Models, Bull. Amer. Meteorol. Soc., 98, 565–588, https://doi.org/10.1175/BAMS-D-15-00268.1, 2017.

Bishop, C. H., Etherton, B. J., and Majumdar, S. J.: Adaptive Sampling with the Ensemble Transform Kalman Filter. Part I: Theoretical Aspects, Mon. Weather Rev., 129, 420–436, https://doi.org/10.1175/1520-0493(2001)129<0420:ASWTET>2.0.CO;2, 2001.

Carrassi, A., Bocquet, M., Bertino, L., and Evensen, G.: Data assimilation in the geosciences: An overview of methods, issues, and perspectives, WIREs Clim. Change, 9, e535, https://doi.org/10.1002/wcc.535, 2018.

Chen, Z., Li, D., Liu, J., Xu, J., and Yang, Q.: Assimilating Observations to Improve Arctic Sea Ice Seasonal Prediction: A Variational Autoencoder Latent Space Particle Filter Approach, J. Geophys. Res.-Oceans, 130, e2024JC022 206, https://doi.org/10.1029/2024JC022206, 2025.

Corradino, C., Jouve, P., La Spina, A., and Del Negro, C.: Monitoring Earth's atmosphere with Sentinel-5 TROPOMI and Artificial Intelligence: Quantifying volcanic SO2 emissions, Remote Sens. Environ., 315, 114 463, https://doi.org/10.1016/j.rse.2024.114463, 2024.

Costa, A., Suzuki, Y., Cerminara, M., Devenish, B., Ongaro, T. E., Herzog, M., Van Eaton, A., Denby, L., Bursik, M., de' Michieli Vitturi, M., Engwell, S., Neri, A., Barsotti, S., Folch, A., Macedonio, G., Girault, F., Carazzo, G., Tait, S., Kaminski, E., Mastin, L., Woodhouse, M., Phillips, J., Hogg, A., Degruyter, W., and Bonadonna, C.: Results of the eruptive column model inter-comparison study, J. Volcanol. Geoth. Res., 326, 2–25, https://doi.org/10.1016/j.jvolgeores.2016.01.017, 2016.

Devenish, B. J., Francis, P. N., Johnson, B. T., Sparks, R. S. J., and Thomson, D. J.: Sensitivity analysis of dispersion modeling of volcanic ash from Eyjafjallajökull in May 2010, J. Geophys. Res.-Atmos., 117, https://doi.org/10.1029/2011JD016782, 2012.

Folch, A. and Mingari, L.: Data Assimilation of Volcanic Clouds: Recent Advances and Implications on Operational Forecasts, pp. 144–156, Special Publications of the International Union of Geodesy and Geophysics, Cambridge University Press, 2023.

Folch, A., Mingari, L., Gutierrez, N., Hanzich, M., Macedonio, G., and Costa, A.: FALL3D-8.0: a computational model for atmospheric transport and deposition of particles, aerosols and radionuclides – Part 1: Model physics and numerics, Geosci. Model Dev., 13, 1431–1458, https://doi.org/10.5194/gmd-13-1431-2020, 2020.

Folch, A., Mingari, L., and Prata, A.: Ensemble-based forecast of volcanic clouds using FALL3D-8.1, Front. Earth Sci., 9, 741 841, https://doi.org/10.3389/feart.2021.741841, 2022.

Folch, A., Abril, C., Afanasiev, M., Amati, G., Bader, M., Badia, R. M., Bayraktar, H. B., Barsotti, S., Basili, R., Bernardi, F., et al.: The EU Center of Excellence for Exascale in Solid Earth (ChEESE): Implementation, results, and roadmap for the second phase, Future Gener. Comp. Sy., pp. 47–61, https://doi.org/j.future.2023.04.006, 2023.

Folch, A., Costa, A., Macedonio, G., Mingari, L., and Pascual Saldaña, H.: FALL3D, https://doi.org/10.5281/zenodo.17737081, 2025.

Fu, G., Heemink, A., Lu, S., Segers, A., Weber, K., and Lin, H.-X.: Model-based aviation advice on distal volcanic ash clouds by assimilating aircraft in situ measurements, Atm. Chem. Phys., 16, 9189–9200, https://doi.org/10.5194/acp-16-9189-2016, 2016.

https://doi.org/10.5194/egusphere-2025-6097
Preprint. Discussion started: 30 December 2025

Fu, G., Prata, F., Lin, H. X., Heemink, A., Segers, A., and Lu, S.: Data assimilation for volcanic ash plumes using a satellite observational operator: a case study on the 2010 Eyjafjallajökull volcanic eruption, Atm. Chem. Phys., 17, 1187–1205, https://doi.org/10.5194/acp-17-1187-2017, 2017.

Gao, Z., Shi, X., Han, B., Wang, H., Jin, X., Maddix, D., Zhu, Y., Li, M., and Wang, Y.: PreDiff: Precipitation Nowcasting with Latent Diffusion Models, https://arxiv.org/abs/2307.10422, 2023.

Goodfellow, I., Bengio, Y., and Courville, A.: Deep Learning, MIT Press, http://www.deeplearningbook.org, 2016.

Govett, M., Bah, B., Bauer, P., Berod, D., Bouchet, V., Corti, S., Davis, C., Duan, Y., Graham, T., Honda, Y., Hines, A., Jean, M., Ishida, J., Lawrence, B., Li, J., Luterbacher, J., Muroi, C., Rowe, K., Schultz, M., Visbeck, M., and Williams, K.: Exascale Computing and Data Handling: Challenges and Opportunities for Weather and Climate Prediction, Bull. Amer. Meteorol. Soc., 105, E2385–E2404, https://doi.org/10.1175/BAMS-D-23-0220.1, 2024.

Kalnay, E.: Atmospheric modeling, data assimilation and predictability, Cambridge university press, 2003.

Kashinath, K., Mustafa, M., Albert, A., Wu, J.-L., Jiang, C., Esmaeilzadeh, S., Azizzadenesheli, K., Wang, R., Chattopadhyay, A., Singh, A., Manepalli, A., Chirila, D., Yu, R., Walters, R., White, B., Xiao, H., Tchelepi, H. A., Marcus, P., Anandkumar, A., Hassanzadeh, P., and Prabhat: Physics-informed machine learning: case studies for weather and climate modelling, Philos. Trans. A Math. Phys. Eng. Sci., 379, 20200 093, https://doi.org/10.1098/rsta.2020.0093, 2021.

Kingma, D. P. and Welling, M.: Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114, 2013.

Leutbecher, M. and Palmer, T.: Ensemble forecasting, J. Comput. Phys., 227, 3515–3539, https://doi.org/10.1016/j.jcp.2007.02.014, predicting weather, climate and extreme events, 2008.

Madankan, R., Pouget, S., Singla, P., Bursik, M., Dehn, J., Jones, M., Patra, A., Pavolonis, M., Pitman, E., Singh, T., and Webley, P.: Computation of probabilistic hazard maps and source parameter estimation for volcanic ash transport and dispersion, J. Comput. Phys., 271, 39–59, https://doi.org/10.1016/j.jcp.2013.11.032, frontiers in Computational Physics, 2014.

Marzocchi, W., Sandri, L., and Furlan, C.: A quantitative model for volcanic hazard assessment, in: Statistics in Volcanology, Geological Society of London, ISBN 9781862392083, https://doi.org/10.1144/IAVCEI001.3, 2006.

McInnes, L., Healy, J., and Melville, J.: UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, https://arxiv.org/abs/1802.03426, 2020.

McKay, M. D., Beckman, R. J., and Conover, W. J.: A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code, Technometrics, 21, 239–245, 1979.

Mingari, L., Folch, A., Prata, A. T., Pardini, F., Macedonio, G., and Costa, A.: Data assimilation of volcanic aerosol observations using FALL3D+PDAF, Atm. Chem. Phys., 22, 1773–1792, https://doi.org/10.5194/acp-22-1773-2022, 2022.

Osores, S., Ruiz, J., Folch, A., and Collini, E.: Volcanic ash forecast using ensemble-based data assimilation: an ensemble transform Kalman filter coupled with the FALL3D-7.2 model (ETKF–FALL3D version 1.0), Geosci. Model Dev., 13, 1–22, https://doi.org/10.5194/gmd-13-1-2020, 2020.

Pardini, F., Corradini, S., Costa, A., Esposti Ongaro, T., Merucci, L., Neri, A., Stelitano, D., et al.: Ensemble-Based Data Assimilation of Volcanic Ash Clouds from Satellite Observations: Application to the 24 December 2018 Mt. Etna Explosive Eruption, Atmosphere, 11, 359, 2020.

Pasmans, I., Chen, Y., Finn, T. S., Bocquet, M., and Carrassi, A.: Ensemble Kalman filter in latent space using a variational autoencoder pair, https://arxiv.org/abs/2502.12987, 2025.

https://doi.org/10.5194/egusphere-2025-6097
Preprint. Discussion started: 30 December 2025

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library, https://arxiv.org/abs/1912.01703, 2019.

Pfeiffer, T., Costa, A., and Macedonio, G.: A model for the numerical simulation of tephra fall deposits, J. Volcanol. Geoth. Res., 140, 273 – 294, https://doi.org/10.1016/j.jvolgeores.2004.09.001, 2005.

Poulidis, A. P. and Iguchi, M.: Model sensitivities in the case of high-resolution Eulerian simulations of local tephra transport and deposition, Atmos. Res., 247, 105 136, https://doi.org/10.1016/j.atmosres.2020.105136, 2021.

Poulidis, A. P., Phillips, J. C., Renfrew, I. A., Barclay, J., Hogg, A., Jenkins, S. F., Robertson, R., and Pyle, D. M.: Meteorological controls on local and regional volcanic ash dispersal, Sci. Rep., 8, 6873, 2018.

Prata, A. T., Mingari, L., Folch, A., Macedonio, G., and Costa, A.: FALL3D-8.0: a computational model for atmospheric transport and deposition of particles, aerosols and radionuclides – Part 2: Model validation, Geosci. Model Dev., 14, 409–436, https://doi.org/10.5194/gmd-14-409-2021, 2021.

Raissi, M., Perdikaris, P., and Karniadakis, G.: Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, J. Comput. Phys., 378, 686–707, https://doi.org/https://doi.org/10.1016/j.jcp.2018.10.045, 2019.

Sandri, L., Tierz, P., and Loughlin, S. C.: From Eruptive Histories to Volcano Monitoring: Probabilistic Eruption Forecasting and Volcanic Hazard Assessment at Varying Temporal and Spatial Scales, pp. 365–397, Springer Nature Switzerland, Cham, ISBN 978-3-031-86841-2, https://doi.org/10.1007/978-3-031-86841-2_12, 2025.

Stefanescu, E. R., Patra, A. K., Bursik, M. I., Madankan, R., Pouget, S., Jones, M., Singla, P., Singh, T., Pitman, E. B., Pavolonis, M., Morton, D., Webley, P., and Dehn, J.: Temporal, probabilistic mapping of ash clouds using wind field stochastic variability and uncertain eruption source parameters: Example of the 14 April 2010 Eyjafjallajökull eruption, J. Adv. Model. Earth Sy., 6, 1173–1184, https://doi.org/https://doi.org/10.1002/2014MS000332, 2014.

Titos, M., Martínez Montesinos, B., Barsotti, S., Sandri, L., Folch, A., Mingari, L., Macedonio, G., and Costa, A.: Long-term hazard assessment of explosive eruptions at Jan Mayen (Norway) and implications for air traffic in the North Atlantic, Nat. Hazards Earth Syst. Sci., 22, 139–163, https://doi.org/10.5194/nhess-22-139-2022, 2022.

Torrisi, F., Corradino, C., Cariello, S., and Del Negro, C.: Enhancing detection of volcanic ash clouds from space with convolutional neural networks, J. Volcanol. Geoth. Res., 448, 108 046, https://doi.org/10.1016/j.jvolgeores.2024.108046, 2024.

Toth, Z. and Kalnay, E.: Ensemble Forecasting at NCEP and the Breeding Method, Mon. Weather Rev., 125, 3297–3319, https://doi.org/10.1175/1520-0493(1997)125<3297:EFANAT>2.0.CO;2, 1997.

Tretiak, D., Mohan, A. T., and Livescu, D.: Physics-Constrained Generative Adversarial Networks for 3D Turbulence, https://arxiv.org/abs/2212.00217, 2022.

Wang, H., Han, J., Fan, W., Zhang, W., and Liu, H.: PhyDA: Physics-Guided Diffusion Models for Data Assimilation in Atmospheric Systems, https://arxiv.org/abs/2505.12882, 2025.

Weisheimer, A., Corti, S., Palmer, T., and Vitart, F.: Addressing model error through atmospheric stochastic physical parametrizations: impact on the coupled ECMWF seasonal forecasting system, Philos. Trans. A Math. Phys. Eng. Sci., 372, 20130 290, https://doi.org/10.1098/rsta.2013.0290, 2014.

Woodhouse, M. J., Hogg, A. J., and Phillips, J. C.: A global sensitivity analysis of the PlumeRise model of volcanic plumes, J. Volcanol. Geoth. Res., 326, 54–76, https://doi.org/10.1016/j.jvolgeores.2016.02.019, 2016.

https://doi.org/10.5194/egusphere-2025-6097
Preprint. Discussion started: 30 December 2025

Wu, H. and Levinson, D.: The ensemble approach to forecasting: A review and synthesis, Transp. Res. Part C-Emerg. Technol., 132, 103 357, https://doi.org/10.1016/j.trc.2021.103357, 2021.

Yang, Z., Wu, J.-L., and Xiao, H.: Enforcing Deterministic Constraints on Generative Adversarial Networks for Emulating Physical Systems, https://arxiv.org/abs/1911.06671, 2020.

Zidikheri, M. J. and Lucas, C.: A Computationally Efficient Ensemble Filtering Scheme for Quantitative Volcanic Ash Forecasts, J. Geophys. Res.-Atmos., 126, e2020JD033 094, https://doi.org/10.1029/2020JD033094, 2021a.

Zidikheri, M. J. and Lucas, C.: Improving Ensemble Volcanic Ash Forecasts by Direct Insertion of Satellite Data and Ensemble Filtering, Atmosphere, 12, https://doi.org/10.3390/atmos12091215, 2021b.