Atmos. Chem. Phys., 26, 15953–15973, 2026
https://doi.org/10.5194/acp-26-15953-2026
© Author(s) 2026. This work is distributed under
the Creative Commons Attribution 4.0 License.

Atmospheric Chemistry and Physics
Open Access
EGU

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

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

[1]Geosciences Barcelona (GEO3BCN-CSIC), Barcelona, Spain
[2]Department of Signal Theory, Telematics and Communications, Research Centre for Information and Communication Technologies (CITIC-UGR), University of Granada, Granada, Spain

Correspondence: Leonardo Mingari (lmingari@csic.es)

Received: 7 December 2025 – Discussion started: 30 December 2025
Revised: 15 September 2026 – Accepted: 1 November 2026 – Published: 11 November 2026

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 forecast uncertainty quantification, data assimilation, and probabilistic hazard assessment, require large ensembles to reduce sampling errors and adequately represent probability distributions. However, simulating and processing large ensembles with atmospheric dispersal models can be computationally demanding, even on high-performance computing clusters. Generative AI offers a complementary approach by representing simulated fields in a lower-dimensional latent space, enabling new samples to be generated at low computational cost and facilitating the manipulation of large ensembles. A Variational AutoEncoder (VAE) is trained on ensembles of physics-based simulations and used to generate new samples. The learned latent-space representation is then used to test two ensemble-based data assimilation methods, demonstrating the potential benefits of performing data assimilation in a lower-dimensional space. As a proof of concept, the data assimilation problem is formulated in two dimensions. 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 %. These results highlight the potential of generative AI to efficiently augment physics-based ensembles and improve forecasting skill through latent-space data assimilation.

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 uncertainties. 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, 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 dispersal 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. Firstly, with the help of ensembles, it has been possible to characterise and quantify forecast uncertainties due to poorly constrained model inputs, 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). Secondly, for volcanic hazard assessment, which is 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 volcanic products (Marzocchi et al., 2006), and a proper probabilistic hazard assessment generally requires a large number of numerical simulations (Madankan et al., 2014; Titos et al., 2022). Finally, forecast errors can be reduced by incorporating observations into numerical models using ensemble-based data assimilation methods (Kalnay, 2002). 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). However, data assimilation remains challenging for volcanic clouds due to the strongly non-Gaussian and highly skewed distribution of ash concentrations, which can limit the applicability of methods based on Gaussian assumptions (Mingari et al., 2023). In turn, particle-based approaches can account for such non-Gaussian distributions, but their application to the high-dimensional model state can suffer from severe particle degeneracy (van Leeuwen, 2009).

Developments in Artificial Intelligence (AI) provide new approaches to modelling volcanic clouds and, specifically, alternative ways of representing complex, high-dimensional data. 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. In particular, Variational Autoencoders (VAEs; Kingma and Welling, 2013) are a class of generative models that 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. The learned latent representation can be used subsequently for the reconstruction and synthesis of new states by sampling from the latent space, enabling the generation of large numbers of new samples (ensemble members) at relatively low computational cost. In addition, such a reduced-dimensional representation may be particularly relevant for ensemble-based data assimilation, as it provides an alternative space in which the probability distribution of the model state can be represented and updated. Several recent studies have demonstrated the benefits of using a low-dimensional latent space for data assimilation purposes (Pasmans et al., 2026). 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, demonstrating that the proposed methodology is a promising solution for non-linear data assimilation in realistic high-dimensional geoscience applications. Alternatively, Pasmans et al. (2026) 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, Pasmans et al. (2026) 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.

This paper explores the potential of generative AI for probabilistic modelling and forecasting of volcanic clouds, proposing a hybrid approach that combines a VAE with physics-based numerical simulations. A conventional VAE is employed to learn a reduced-dimensional latent representation of the statistical structure of an ensemble of simulations. As an initial proof of concept, an ensemble of two-dimensional mass loading, i.e. vertical column mass per unit area of airborne ash, was used as training data for a convolutional neural network. The training, validation, and test datasets were generated using the VATD model FALL3D (Folch et al., 2020, 2025; Prata et al., 2021). Two numerical experiments are presented. In the first synthetic experiment, the VAE is used to generate thousands of new samples from a small ensemble of numerical simulations, and its ability to reproduce the statistical characteristics of a larger reference ensemble is evaluated. In the second experiment, satellite retrievals from the 2018 Etna eruption are used to evaluate two different latent-space data assimilation

filters, the Ensemble Transform Kalman Filter (ETKF) and a regularised bootstrap Particle Filter (PF). The manuscript is organised as follows. A brief theoretical background on VAEs and a description of the neural network architecture are presented in Sect. 2. The ensemble-based data assimilation methods employed in this work are described in Sect. 3. Section 4 describes the numerical experiments, including the FALL3D numerical model, model configurations used, and performance evaluation metrics. The results are presented and discussed in Sect. 5 and, finally, conclusions are drawn in Sect. 6, together with a discussion on the implications and potential applications of the proposed approach.

2 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. By sampling 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 (Kingma and Welling, 2013). The scheme in Fig. 1 shows a typical architecture of a VAE composed of convolutional network 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 $\boldsymbol{x}$.

The training of the VAE is guided by a loss function ($\mathcal{L}_{\text{VAE}}$) 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 $p(\boldsymbol{z})$:

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

where $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ is the encoder approximate posterior and $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ the decoder likelihood. The first term ($\mathcal{L}_{\text{R}}$, reconstruction loss) encourages reconstructions to be accurate and, in this work, the loss function is assumed to be the Mean Squared Error (MSE) between the input ($\boldsymbol{x}$) and its reconstructed output ($\boldsymbol{x}'$):

$$\mathcal{L}_{\text{R}} = \frac{1}{N}\sum_{i=1}^{N}(x_i - x_i')^2 \quad (2)$$

being N the dimension of the original space. The second term ($\mathcal{L}_{\text{KL}}$) defines the loss function in terms of the Kullback-Leibler divergence (D_{KL}), which measures the distance 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, $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{0}, \mathbb{I})$, leading to:

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

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)^\intercal$ is the vector of means, $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_K^2)$ is the vector of variances, and K the dimension of the latent space where vectors $\boldsymbol{z}$ live. In addition, this work uses the so-called β-VAE, a modification of the standard VAE that introduces a hyper-parameter β to weight the $\mathcal{L}_{\text{KL}}$ term in the global VAE loss function:

$$\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{R}} + \beta\,\mathcal{L}_{\text{KL}} \quad (4)$$

Larger values of β increasingly regularize the latent space towards the standard Gaussian prior $\mathcal{N}(\boldsymbol{0}, \mathbb{I})$, potentially reducing the amount of information encoded in the latent representation. In summary, the VAE training step optimises the encoder and the decoder by minimising the VAE loss function Eq. (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. Details on VAE network architecture and hyper-parameter values are given in Sect. 4.4.

3 Latent-space data assimilation

Data Assimilation (DA) techniques aim to obtain an optimal state of a dynamical system by combining model forecasts with observations, and are widely used by operational centres to forecast geophysical systems (Carrassi et al., 2018). However, the application of DA to geophysical systems is challenging because their system states are typically represented by very high-dimensional vectors. As the dimensionality of a system increases, DA methods become increasingly difficult to apply due to the complexity of representing the system state and its associated uncertainties. Ensemble-based DA methods partially address the issue of high-dimensionality by representing the uncertainty of the system state through a finite ensemble of model realizations. The key underpinning concept is that the ensemble can approximate the relevant error statistics, making ensemble-based methods practical for high-dimensional and non-linear geophysical systems. Still, in many geoscientific applications, the computational cost of the physical models constrains the ensemble sizes, increasing sampling errors in the estimation of ensemble statistics and potentially leading to suboptimal analyses.

These limitations are particularly relevant for atmospheric applications, where the system state may comprise millions of degrees of freedom, while non-linear dynamics and non-Gaussian error distributions can limit the applicability of traditional DA approaches. For example, many variational and Kalman filter methods rely on assumptions of approximately Gaussian error statistics, leading to inaccurate state

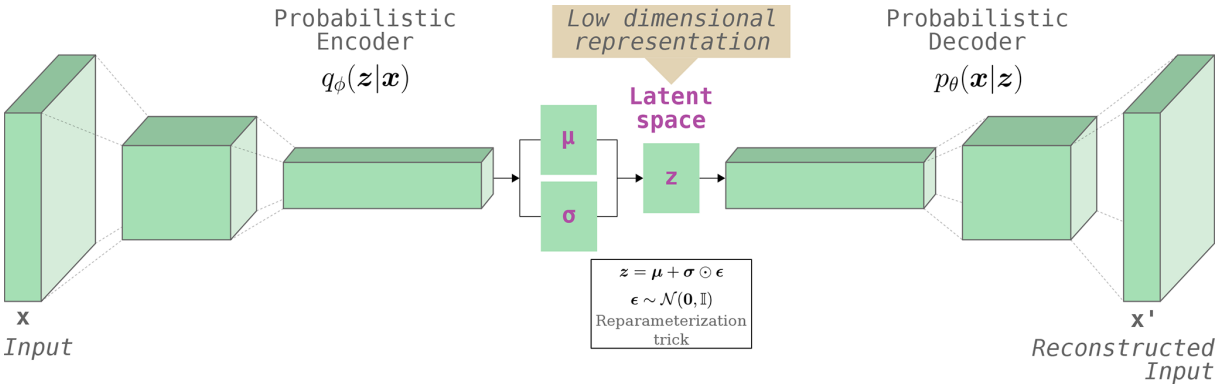


Figure 1. The basic scheme of a VAE composed of two-dimensional convolutional layers (Conv2D). The encoder compresses an input vector $\boldsymbol{x}$ into the low-dimensional latent space. The decoder attempts to reconstruct the original data from a lower-dimensional vector $\boldsymbol{z}$ sampled from the latent space. 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.

estimates when the underlying distributions are strongly non-Gaussian. Particularly challenging for DA are variables that are lower-bounded (e.g. cannot take negative values) and poorly described by Gaussian distributions, such as water-vapour mixing ratios (Kliewer et al., 2016), rainfall (Husak et al., 2007), aerosol concentrations (O'Neill et al., 2000), and volcanic ash concentrations (Mingari et al., 2023), often exhibiting strongly skewed and near-zero distributions. Particle filters circumvent the Gaussian assumption by representing the state distribution with a set of weighted samples (particles) and updating their weights according to the observation likelihood. However, as the number of state variables grows, particle weights can collapse onto one or a few particles, leading to severe particle degeneracy.

The low-dimensional manifold learned by the VAE can be exploited to perform DA in a compressed representation of the model state, expressed in terms of a relatively small number of latent variables while retaining the main spatial structures and variability of the original fields. Thus, rather than applying DA directly to the high-dimensional ash concentration fields, the analysis is performed in the latent space, reducing the dimensionality of the problem and then using the VAE decoder to map the resulting states back to the physical space. The application of ensemble-based DA methods in the VAE latent space offers several advantages. First of all, the low dimensionality of the latent representation substantially reduces the computational cost of the DA analysis, making it possible to use larger ensembles than in the original physical space. In turn, larger ensembles can potentially reduce sampling errors in the estimation of ensemble statistics. On the other hand, the compact latent representation facilitates the estimation of these statistics, while the approximately Gaussian distribution of the latent variables makes the latent space particularly suitable for ensemble-based DA methods based on Gaussian assumptions. In this work, we investigate two ensemble-based DA approaches operating in the VAE latent space: the Ensemble Transform Kalman Filter (ETKF) and the regularized bootstrap Particle Filter (PF), briefly described below.

3.1 Ensemble formulation

Let $\{z_i\}$ denote a prior ensemble of N_E latent-space vectors, with $z_i \in \mathbb{R}^K$, and let $\boldsymbol{y}_\mathrm{obs} \in \mathbb{R}^M$ denote the vector of M observations. The latent-space ensemble is mapped to the physical space through the VAE decoder D,

$$\boldsymbol{x}_i = D(z_i) \tag{5}$$

with $\boldsymbol{x}_i \in \mathbb{R}^N$. The corresponding observation-space ensemble is then obtained by applying the observation operator,

$$\boldsymbol{y}_i = H(\boldsymbol{x}_i) = H(D(z_i)), \tag{6}$$

where H is the operator mapping the physical state to the observation space. Thus, although the analysis is performed in the K-dimensional latent space, the observations are related to the latent variables through the composition of the VAE decoder and the observation operator.

The prior ensemble mean in latent space is defined as

$$\bar{z} = \frac{1}{N_\mathrm{E}} \sum_{i=1}^{N_\mathrm{E}} z_i \tag{7}$$

We define the background anomaly matrix $\mathbf{Z}' \in \mathbb{R}^{K \times N_\mathrm{E}}$, whose ith column represents the deviation of the corresponding ensemble member from the ensemble mean, $z_i - \bar{z}$. Similarly, we define the observation-space anomaly matrix $\mathbf{Y}' \in \mathbb{R}^{M \times N_E}$, whose columns represent the deviations of the observation-space ensemble members from the corresponding ensemble mean, $\boldsymbol{y}_i - \bar{\boldsymbol{y}}$.

3.2 Ensemble transform Kalman filter

The Ensemble transform Kalman filter (ETKF), introduced by Bishop et al. (2001), is employed in this work following the formulation adopted by Hunt et al. (2007). The analysis ensemble is obtained by updating both the prior ensemble mean ($\bar{z} \rightarrow \bar{z}_a$) and the prior ensemble anomalies ($\mathbf{Z}' \rightarrow \mathbf{Z}'_a$). The analysis ensemble mean $\bar{z}_\mathrm{a}$ is expressed as a linear combination of the background ensemble anomalies,

$$\bar{z}_\mathrm{a} = \bar{z} + \mathbf{Z}'\,\boldsymbol{w}, \tag{8}$$

where $\boldsymbol{w} \in \mathbb{R}^{N_\mathrm{E}}$ is a vector of weights to be determined. In turn, the analysis anomalies $\mathbf{Z}'_\mathrm{a}$ are obtained by applying a transformation matrix $\mathbf{W}$ to the background anomalies,

$$\mathbf{Z}'_\mathrm{a} = \mathbf{Z}'\,\mathbf{W}. \tag{9}$$

The weight vector $\boldsymbol{w}$ in Eq. (8) is obtained by solving the linear system

$$\mathbf{A}\boldsymbol{w} = \mathbf{Y}'^{\mathsf{T}}\mathbf{R}^{-1}\boldsymbol{d}, \tag{10}$$

where $\mathbf{R}$ is the observation-error covariance matrix, $\boldsymbol{d} = \boldsymbol{y}_\mathrm{obs} - \bar{\boldsymbol{y}}$ is the innovation vector, and the matrix $\mathbf{A}$ is defined as

$$\mathbf{A} = (k-1)\mathbf{I} + \mathbf{Y}'^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{Y}'. \tag{11}$$

The analysis covariance in the latent space is given by

$$\mathbf{P}^\mathrm{a} = \mathbf{Z}'\mathbf{A}^{-1}\mathbf{Z}'^{\mathsf{T}} \tag{12}$$

In order to obtain an ensemble with this covariance, the transformation matrix $\mathbf{W}$ in Eq. (9) is constructed as a symmetric square root of $(N_\mathrm{E}-1)\mathbf{A}^{-1}$. Specifically, using the eigendecomposition $\mathbf{A} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathsf{T}}$ the transformation matrix $\mathbf{W}$ is given by

$$\mathbf{W} = \sqrt{N_\mathrm{E}-1}\,\mathbf{V}\boldsymbol{\Lambda}^{-1/2}\mathbf{V}^{\mathsf{T}} \tag{13}$$

This transformation ensures that the sample covariance of the analysis ensemble is consistent with $\mathbf{P}^\mathrm{a}$. The posterior latent ensemble members are obtained from Eqs. (8) and (9) by adding the transformed anomalies to the analysis mean. Finally, the resulting set of vectors z_a are then be passed through the VAE decoder to obtain the corresponding physical-space ensemble resulting from the analysis step.

3.3 Regularized bootstrap particle filter

Particle filtering (Doucet et al., 2001) has been successfully applied to low-dimensional models, e.g. in non-linear tracking problems (Arulampalam et al., 2002), but direct application of the basic PF to high-dimensional systems is prone to weight collapse, as discussed previously. As a result, several variants have been proposed to enable the application of particle filters in geophysical systems (van Leeuwen, 2009). Since the dimensionality of the latent space is substantially lower than that of the original physical space, we explore whether particle filtering becomes a viable option for data assimilation in this reduced representation. In this work, a basic bootstrap particle filter (Gordon et al., 1993) with Gaussian kernel regularization (Pham, 2001) is tested for latent-space data assimilation. The particle weights are computed from the observation likelihood,

$$w_i \propto p\left(\boldsymbol{y}_\mathrm{obs} \mid \boldsymbol{y}_i\right), \tag{14}$$

and normalized to sum to one. The particles are then resampled according to their weights using systematic resampling (Kitagawa, 1996), producing an equally weighted ensemble of size N_E. When the particle weights are highly concentrated on a small number of particles, resampling can substantially reduce the diversity of the ensemble. To restore particle diversity, the resampled particles, z_i^r, are subsequently perturbed using a Gaussian kernel,

$$z_i^\mathrm{a} = z_i^\mathrm{r} + \boldsymbol{\epsilon}_i, \qquad \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\mathrm{reg}), \tag{15}$$

where $\boldsymbol{\Sigma}_\mathrm{reg} = h^2\mathbf{P}$, with $\mathbf{P}$ the sample covariance of the prior latent ensemble and h a prescribed regularization factor. This kernel perturbation provides a simple regularization aimed at increasing the diversity among the resampled particles.

4 Numerical experiments and evaluation

Two types of numerical experiments were performed: (i) a synthetic experiment (Mingari, 2026b) to characterize the VAE-generated ensemble and assess its ability to reproduce the statistics of a physics-based reference ensemble; and (ii) a second experiment (Mingari, 2026a) considering a real-case application aimed at improving the modelling of volcanic clouds through data assimilation in the latent space. The code and the material required to reproduce both experiments are available in the Supplement.

4.1 Physics-based atmospheric dispersal model

The latest release (version 9.1) of the open-source code FALL3D (Folch et al., 2025) was used to produce an ensemble of numerical simulations for the two experiments. The resulting datasets were then used for VAE training, testing and evaluation. 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 FALL3D model

has been deployed and optimised on several EuroHPC supercomputing systems: 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. The configuration of the FALL3D model for the two experiments is summarised in Table 1.

4.2 Synthetic experiment

To define the first experiment, we used model data from the recent EU-MODEX Tenerife 2025 exercise (European Commission, 2025), organised under the European Civil Protection Mechanism. This exercise represented the first large-scale volcanic eruption drill in Spain to practice the response to a potential volcanic eruption from Garachico volcano (Tenerife, Canary Islands) and test the evacuation, the social assistance and communication protocols 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. Here we use the resulting simulations to characterize the statistical properties of an augmented ensemble generated by the VAE in a controlled experiment. In this context, the drill scenario refers to a fictitious eruption defined by given emission source parameters (eruption column height, mass eruption rate, and timing), rather than being constrained by real-time observations as would happen during a real event. 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.

Three ensembles were simulated with FALL3D to produce the training, validation, and test datasets. The training ensemble, consisting of 256 members, was used to train the VAE and the validation ensemble, consisting of 64 members, was used during the training step to monitor the performance and tuning of the neural network hyper-parameters. Finally, the larger test ensemble, consisting of 2048 members, 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 supercomputing system, hosted at the Finish CSC. In particular, the execution of the test ensemble required substantial HPC resources and was executed on the LUMI-G partition allocating 256 computing nodes and a total of 2048 GPUs (one per ensemble member). In all cases, a 24 h time window forecast was performed assuming a continuous ash emission starting at 09:00 UTC (see Table 1), a local-scale computational domain with a horizontal resolution of 0.1° and $120 \times 100 \times 25$ grid cells, and the FALL3D dispersal model was driven by the Global Forecast System (GFS) weather forecast from the National Centers for Environmental Prediction (NCEP). Each ensemble was simulated by perturbing Eruption Source Parameters (ESPs) and wind components, which are widely recognized as among the most sensitive parameters/inputs for volcanic ash transport models (Devenish et al., 2012; Poulidis et al., 2018). The column height was sampled around a central value of 9 km above the vent level. Further details on the perturbed parameters and sampling ranges are provided in the Supplement.

4.3 Etna experiment

The second experiment explores the feasibility of the latent-space DA using a VAE-generated prior ensemble for the 24 December 2018 Etna eruption (Corradini et al., 2020; Pardini et al., 2020). In this case, the FALL3D model was driven by the Copernicus regional reanalysis for Europe (CERRA; Schimanke et al., 2021; Ridal et al., 2024), with 106 vertical model levels, and initialized at 10:00 UTC. An 8 h time window forecast was performed using a 128-member ensemble (model prior). An additional 48-member ensemble was simulated to be used as a validation dataset during the training process (see Table 1). Each ensemble was simulated by perturbing the eruptive column-top height and wind components by 20 % (see Supplement for further details).

The 128-member model prior was used to train a new VAE, which was subsequently used to sample a 4096-members ensemble (VAE prior) for the analysis step. As a proof-of-concept, we used two-dimensional ash column-mass fields rather than the full three-dimensional concentration field, both to reduce the dimensionality of the problem and to maintain a consistent neural-network architecture throughout the study. The input training data, defined on a 256×128 grid ($N = 131\,072$), was encoded into 32-dimensional latent space vectors. Before normalization, a $\log(1+x)$ transformation was applied to the ash column mass ($\mathbf{x}$) to reduce the strong positive skewness of its distribution, with $\mathbf{x}$ expressed in mg m^{-2}.

A single analysis step at 18:00 UTC was performed by assimilating Spinning Enhanced Visible and Infra-Red Imager (SEVIRI) volcanic ash retrievals (Corradini et al., 2020, 2021) considering two ensemble-based approaches, i.e., the ETKF (Sect. 3.2) and the PF (Sect. 3.3). Only positive retrievals were considered, resulting in 2678 valid ob-

Table 1. FALL3D model configurations used in the numerical experiments.

Parameter	Synthetic experiment	Etna experiment
Start time	26 Sep 2025 09:00 UTC	24 Dec 2018 10:00 UTC
Run time	24 h	8 h
Horizontal grid resolution	$0.1 \times 0.1^\circ$	$0.02 \times 0.03^\circ$
Grid size	120×100	256×128
Vertical levels	25	40
Grain size distribution	bi-Gaussian	bi-Gaussian
Emission source	Suzuki source[a]	Suzuki source[a]
Mass emission rate	Estimated[b]	Custom
Ensemble size	256 (train), 64 (validation), 2048 (test)	128 (train), 48 (validation)

[a] Suzuki source formulation following Pfeiffer et al. (2005); [b] Mass emission rate computed following Woodhouse et al. (2016).

servations of ash column mass. A diagonal observation-error covariance matrix was assumed, with the observation-error standard deviation set to 50 % of the retrieved ash column mass, subject to a minimum value of $100\,\mathrm{mg\,m^{-2}}$, representative of a typical instrument detection limit.

4.4 VAE configuration

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 2. The resulting number of trainable parameters was 698 849 and 1 783 489 for the synthetic and the Etna experiments respectively. The neural network was trained using the Adam optimizer during 400 epochs using mini-batches of 16 samples. The reconstruction loss and KL divergence (averaged over batches) were monitored at each epoch for both the training and validation datasets to prevent overfitting. Training was performed using the simulated ensembles of two-dimensional volcanic ash column-mass fields at a single forecast time. As outlined in Fig. 1, the encoder network maps the input data $\boldsymbol{x}$ through successive convolutional layers, followed by a flattening operation to two distinct latent space parameter vectors $\boldsymbol{\mu}$ (mean) and $\log \boldsymbol{\sigma}^2$ (log-variance), with dimensions $K = 16$ and $K = 32$ for the synthetic and Etna experiments, respectively. The decoder then reconstructs the input from the sampled latent vector $\boldsymbol{z}$ using a fully connected layer followed by successive upsampling and convolutional operations, restoring the original spatial dimensions. For the synthetic experiment, transposed convolutional layers are used for upsampling, whereas bilinear interpolation followed by convolutional layers is used in the Etna experiment to reduce checkerboard artifacts.

4.5 Evaluation metrics

4.5.1 Standard performance metrics

The performance of the generated ensembles is evaluated in terms of a set of statistical metrics. Let $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$ denote the reference and estimated quantities, respectively, with N denoting the total number of values being compared. The root mean square error (RMSE) is defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(x_j - \hat{x}_j)^2}. \tag{16}$$

The root mean square logarithmic error (RMSLE) is computed analogously after applying the logarithmic transformation $\log(1+x)$, providing a metric more sensitive to errors across multiple orders of magnitude. Similarly, the Mean Bias Error (MBE) is computed as

$$\mathrm{MBE} = \frac{1}{N}\sum_{j=1}^{N} x_j - \hat{x}_j \tag{17}$$

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

$$\mathrm{MAPE} = 100 \times \sum_{j=1}^{N}\left|\frac{x_j - \hat{x}_j}{\hat{x}_j}\right| \tag{18}$$

Finally, the Pearson correlation coefficient measures the linear correlation between the estimated and reference fields.

4.5.2 Distributional assessment

In addition to the standard performance metrics introduced above, the ability of the VAE to reproduce the statistical properties of the target distribution is further assessed using distributional and spectral metrics. The Jensen-Shannon (JS) divergence is a measure of the similarity between two probability distributions. Specifically, for two distributions

Table 2. Hyper-parameters used for the VAE training in the synthetic and Etna experiments.

Hyperparameter	Synthetic	Etna
Latent dimension K	16	32
Input data size	120×100	256×128
Number of epochs	400	400
Batch size	16	16
Convolutional layers	6	8
Output channels	16, 32, 64	16, 32, 64, 128
Kernel size	3×3	3×3
Stride	1×1	2×2
Activation function	ReLU	ReLU/Softplus
Learning rate	1E-3	1E-3
Optimizer	Adam	Adam
Normalization	Percentile-based scaling	Logarithmic scaling

$P = \{p_i\}$ and $Q = \{q_i\}$, the JS divergence is a symmetrized version of the KL divergence:

$$\mathrm{JSD}(P \parallel Q) = \frac{1}{2} D_{\mathrm{KL}}(P \parallel M) + \frac{1}{2} D_{\mathrm{KL}}(Q \parallel M) \tag{19}$$

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):

$$D_{\mathrm{KL}}(P \parallel Q) = \sum p_i \log \frac{p_i}{q_i} \tag{20}$$

The discrete distributions in Eq. (20) are estimated from histograms in this work.

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

$$P(k) = \frac{1}{N_k} \sum \mathrm{PSD}(k_x, k_y) \tag{21}$$

where $k = \sqrt{k_x^2 + k_y^2}$ is the magnitude of the wavevector, $\mathrm{PSD}(k_x, k_y)$ 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.

4.5.3 Categorical metrics

Categorical metrics were also used to assess binary plume detection. A positive detection is defined when the ash column mass exceeds a given threshold, set to $100\,\mathrm{mg\,m^{-2}}$ in this work for demonstration purposes, which is a typical value corresponding to the satellite detection limit for ash clouds. The probability of detection (POD) measures the fraction of observed plume grid points that are correctly identified,

$$\mathrm{POD} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{22}$$

where TP and FN denote true positives and false negatives, respectively. The Jaccard index measures the spatial overlap between the plume masks,

$$J = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \tag{23}$$

where FP denotes false positives. In summary, the probability of detection (POD) and Jaccard index were used to quantify the detection skill and spatial overlap, respectively.

5 Results and discussion

This section presents the results organized as follows. First, in Sect. 5.1, the synthetic experiment is used to assess the consistency of the VAE training, as well as to perform a sensitivity study to determine the most appropriate configuration. Second, Sect. 5.2 evaluates the capability of the VAE-generated ensemble to reproduce the statistical properties of a larger physics-based ensemble. To this end, a test dataset comprising 2048 FALL3D simulations was generated for the synthetic experiment, serving as a reference for comparison. Third, Sect. 5.3 illustrates a real-case application of this methodology using the Etna experiment: the latent-space assimilation of satellite observations. Finally, the implications of these results and the main limitations of the proposed approach are discussed in Sect. 5.4.

5.1 Training sensitivity analysis

A sensitivity analysis was conducted to evaluate the performance of different neural network configurations trained with the synthetic experiment datasets. Multiple VAE configurations were trained by exploring a range of values for

two key hyper-parameters: the latent space dimension K and the factor β in Eq. (4), while fixing the remaining hyper-parameters with the values specified in Table 2. For each trained VAE, an ensemble was generated by independently sampling latent vectors from the standard normal distribution and decoding them to recover the corresponding physical-space variables. In this section, each generated ensemble consisted of 2048 members.

The performance was evaluated in terms of the RMSE using Eq. (16), computed between the ensemble mean of the test dataset (taken as the reference) and the means of the VAE-generated ensembles. To account for the stochastic variability of the VAE generation, 20 ensembles were generated for each set of hyper-parameters tested. The resulting RMSE values are shown as box plots in Fig. 2. 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. 2, 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. 2a). 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. 2 shows also 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 further supported by the structure of the latent space shown in Fig. 3. The nearly diagonal covariance matrix (Fig. 3a) indicates weak correlation between the latent components, consistent with a well-regularized latent representation. Figure 3b shows a two-dimensional visualization

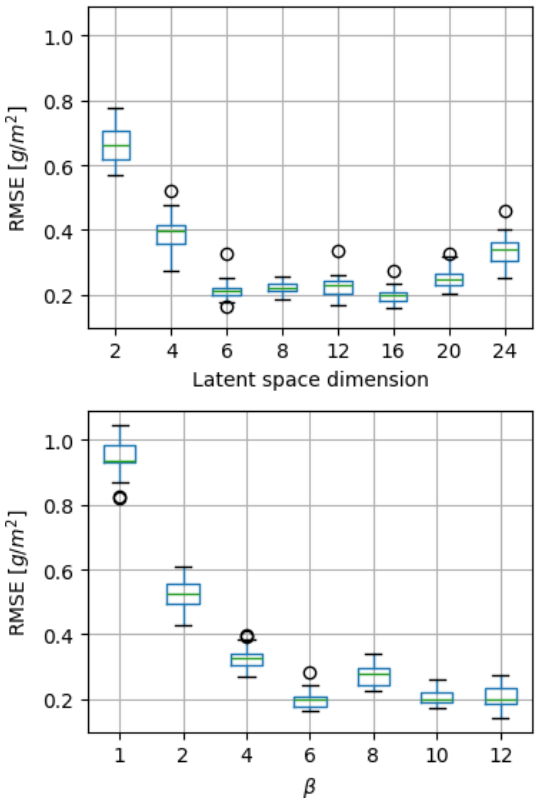


Figure 2. VAE sensitivity analysis for different combinations of two hyper-parameters: the latent space dimension K (top) and the β parameter (bottom), defined in Eq. (4). The performance is measured in terms of the Root-Mean-Square Error (RMSE), computed from the ensemble means. The boxes show the interquartile range, the horizontal lines indicate the median, and the circles represent outliers.

of the latent space based on the Uniform Manifold Approximation and Projection (UMAP) algorithm (McInnes et al., 2020), a technique for non-linear dimensionality reduction specifically designed to preserve the local structure of the data. The UMAP projection shows that the test dataset (circular markers) occupies the same region of latent space than the training data (plus-shaped markers), indicating that the learned representation generalizes to unseen data, rather than only organizing the training samples. 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. 3b 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 a clustering-like structure. These results indicate that the network was trained correctly and behaves as expected. Based on this sensitivity analysis, a latent

dimension of $K = 16$ and $\beta = 6$ were adopted for the VAE configuration used in the next section.

5.2 Evaluation of the VAE-generated ensemble

Once the VAE neural network is properly trained, new samples (ensemble members) can be generated at negligible computational cost. This section explores whether these VAE-generated samples can reproduce the statistical characteristics of an independent and larger ensemble of numerical simulations (test dataset from the synthetic experiment), even when the new samples have not been obtained by explicitly solving the underlying physical equations. For illustrative purposes, Fig. 4 shows a few VAE-generated samples of two-dimensional ash column mass. The variability in the ensemble is a consequence of the different eruptive source parameters (e.g., eruptive column height) and meteorological conditions used to simulate the training dataset. For example, samples in Fig. 4c,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. 4h). In most cases, the volcanic plume is advected eastwards from the vent (16.76° W, 28.27° N). 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.

5.2.1 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 and a few metrics are reported. Figure 5 compares the ensemble means of the training ensemble (Fig. 5a) and VAE-generated ensemble comprising 5000 members (Fig. 5b). Ideally, VAE should be able to closely approximate the test dataset (reference) to achieve good performances. Deviations from the reference test dataset are quantified by bias maps. In relative terms, the bias is small for both the training dataset (Fig. 5c) and VAE (Fig. 5d), although the training dataset provides a slightly better estimate of the ensemble mean in general.

A summary of the metrics for the ensemble mean and spread is shown in Table 3. Metrics are computed for multiple VAE-generated ensembles composed of 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 test dataset (2048 members) is taken as a reference in order to evaluate each of the metrics described in Sect. 4.5. Note that the VAE errors measured in terms of the RMSE are twice as high as those from the training dataset (256 members). In relative terms, this represents an error of 1.78 % (training) and 3.4 % (VAE) according to the MAPE evalu-

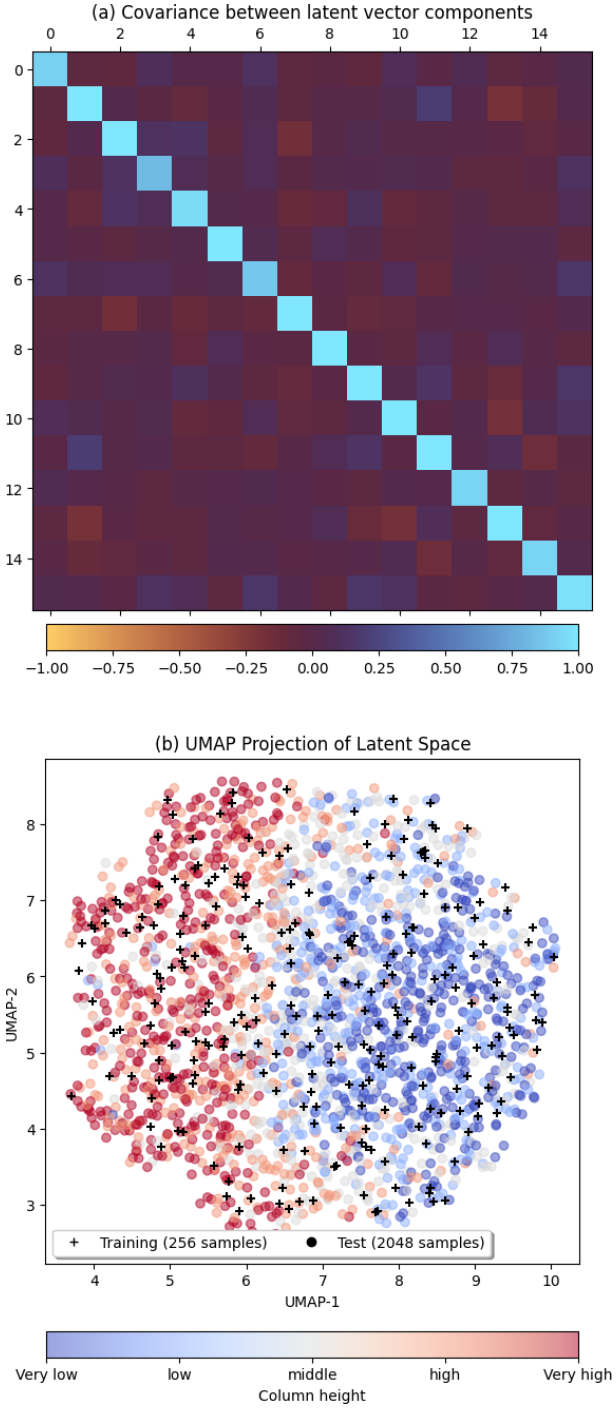


Figure 3. Structure of the latent space for $K = 16$ and $\beta = 6$. The independence between the distributions of the latent vector components can be confirmed from the covariance matrix **(a)**. The test dataset is encoded in the same region of the latent space covered by the training dataset, as shown in the UMAP two-dimensional visualization of the latent space **(b)**, confirming that the learned representation generalizes well to unseen data. See text for details.

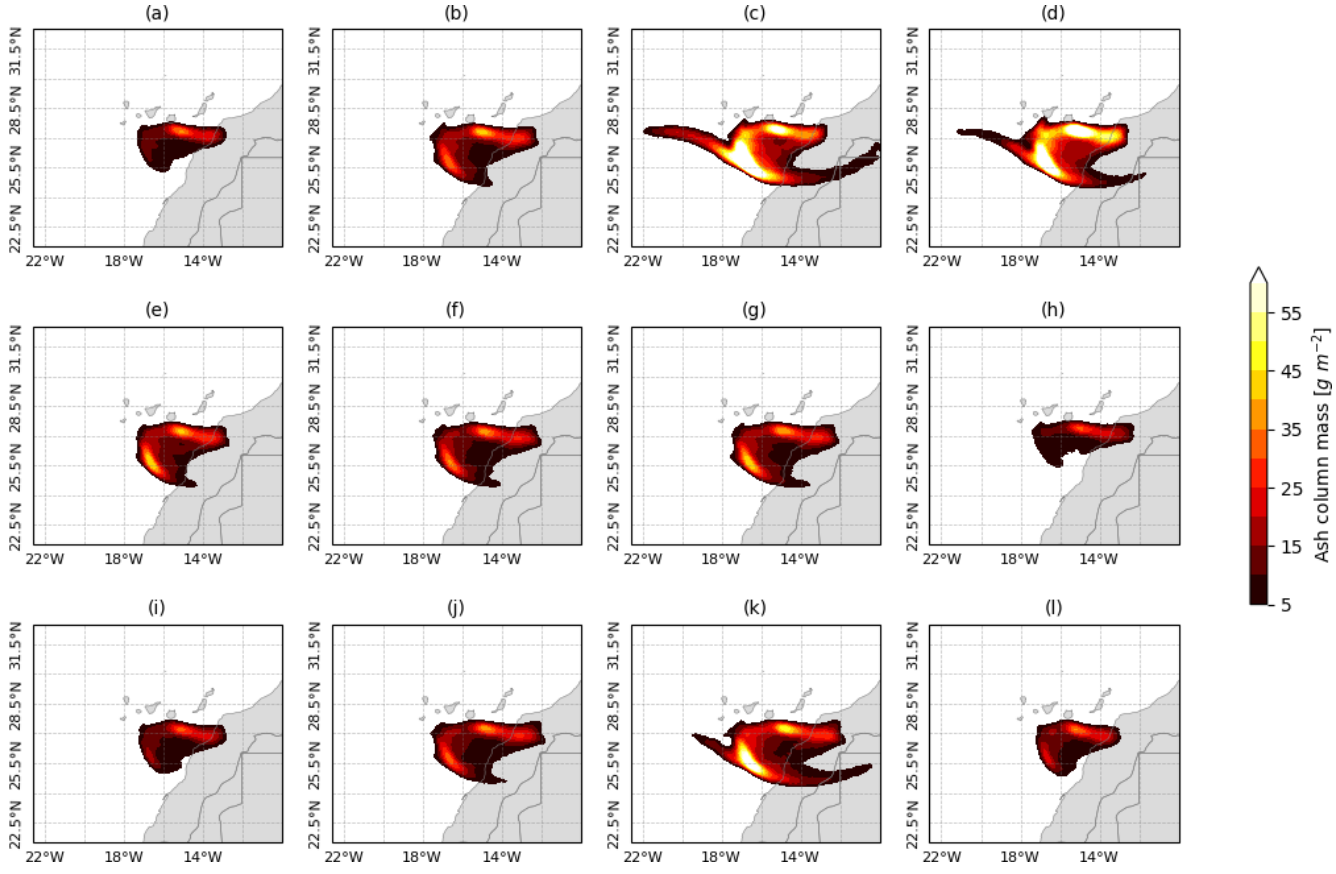


Figure 4. The neural network was trained to generate samples of two-dimensional ash column mass, i.e. the total mass of volcanic ash in a vertical column. In this example, a VAE-generated ensemble composed of 12 samples is shown.

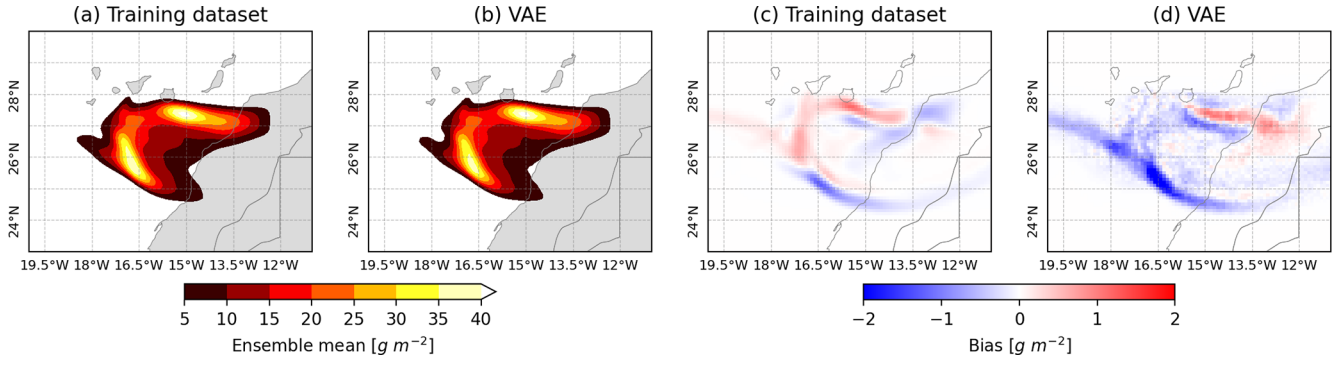


Figure 5. Ensemble mean maps obtained from the training dataset **(a)** and a VAE-generated ensemble with 5000 samples **(b)**. The bias maps for the training **(c)** and VAE **(d)** ensemble means are computed by subtracting the reference (test dataset) corresponding results.

ated from the ensemble mean. In order to avoid divisions by zero and numerical instabilities, the MAPE was actually computed considering only the regions where the test dataset is above 5 g m^{-2} (i.e. the range shown in Fig. 5a,b). We can conclude that the VAE-generated ensemble is properly capturing the first (mean) and second moments (variance) of the distribution, although the training estimations for the ensemble mean and spread are still slightly better. However, the next section demonstrates how, when larger ensembles are generated, the VAE tends to perform better on metrics that are more sensitive to sampling errors.

5.2.2 Probability distributions

An important aspect in hazard and impact assessment is to evaluate exceedance probabilities, i.e. the probability that

Table 3. A summary of some performance metrics evaluated using the test dataset (2048 members) as the reference. The VAE metrics were averaged over 20 ensembles generated to account for stochastic variability. The VAE uncertainty range is given by the standard deviation.

Metric	Variable	Training dataset (256 members)	VAE (5000 members)
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

Values are reported in g m^{-2}, except MAPE, which is given as a percentage (%).

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 6 presents the exceedance probabilities for three different thresholds, demonstrating that a 5000-member VAE-generated ensemble can produce accurate probability distributions as well as realistic samples. In order to examine whether the sampled ensemble provides a good approximation to the reference distribution, a few histograms are shown in Fig. 7 at four selected locations, 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 ($N_E = 256$) 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. 7a).

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 Eq. (19). 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. 8, 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 ($N_E \gtrsim 256$), reaching an asymptotic value of approximately JSD ~ 0.042, which represents a 2.6× improvement over the corresponding training metric (JSD = 0.11).

5.2.3 Isotropic power spectrum

The isotropic power spectrum, i.e. the radially averaged two-dimensional Power Spectral Density (PSD), was computed using Eq. (21). Figure 9 presents the spectra for the test dataset and VAE-generated samples, averaged over the $N_E = 5000$ realizations of each ensemble. Unlike the classical turbulent energy-cascade spectrum, which is often characterized by a power-law dependence proportional to $k^{-5/3}$, the test dataset exhibits an approximately exponential decay (solid red line). This behaviour is likely related to the diffusive processes represented in the FALL3D dispersal model. The results for the transposed-convolution decoder used in the synthetic experiment are represented by green triangles in Fig. 9. The generative model reproduces the large-scale spatial structure reasonably well, but exhibits excess high-wavenumber spectral power at scales smaller than approximately 3 Δx, where Δx denotes the grid spacing, indicating the introduction of artificial small-scale noise. The flat high-wavenumber tail in the two-dimensional power spectrum may be related to the convolutional kernel size used in this work (i.e., 3×3) and to limitations of the decoder upsampling operation. The decoder architecture used in the synthetic experiment is based on transposed convolution layers, which are known to produce chequerboard artefacts. A similar pattern is visible in the bias map of the ensemble-mean field (Fig. 5d), where a noisy pattern with relatively low intensity and high-frequency structure can be identified. 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.

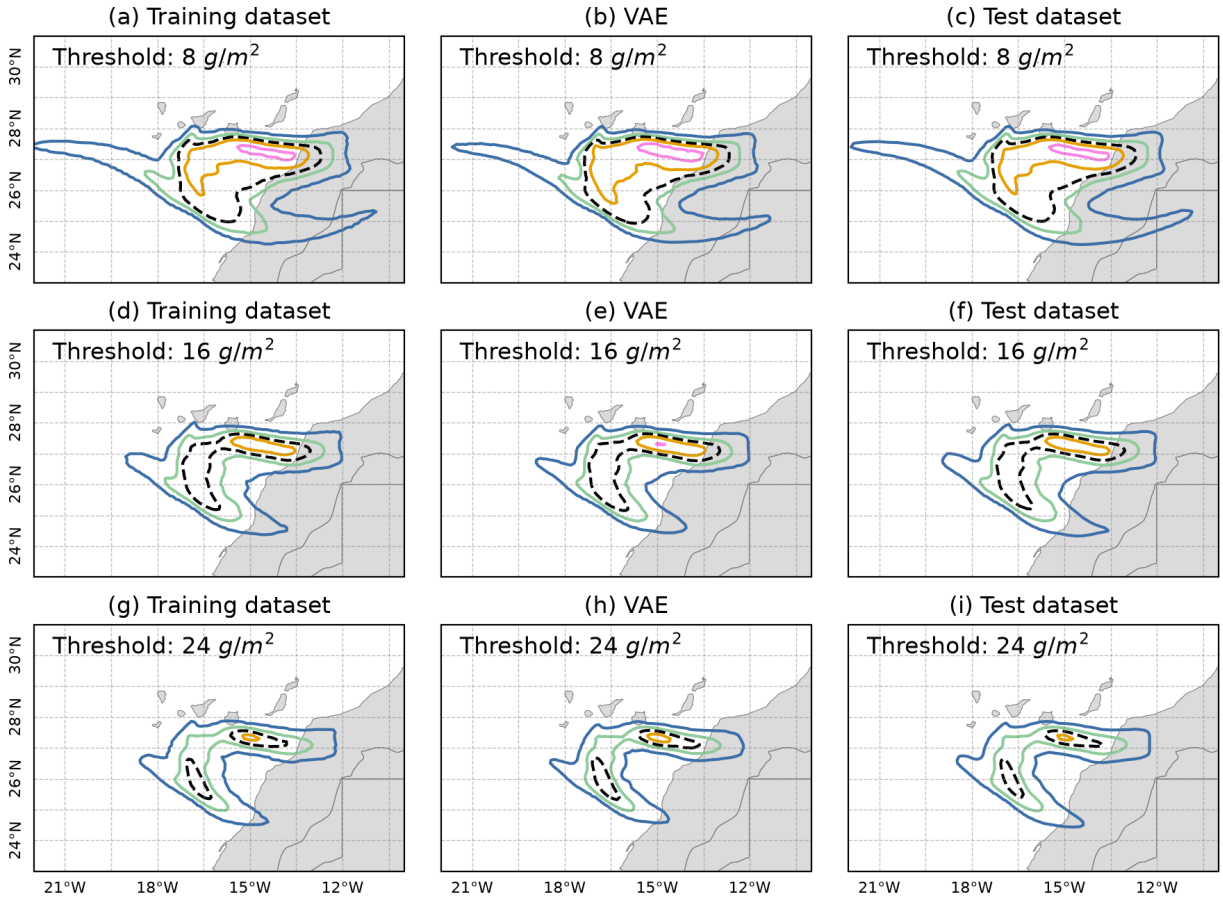


Figure 6. Exceedance probabilities for three ash column mass thresholds (8, 16 and 24 g m^{-2} from top to bottom). The probability contours indicate probability levels of 2 % (blue), 25 % (green), 50 % (dashed black), 75 % (orange), and 98 % (magenta). Results for the training dataset (left column) and a VAE-generated ensemble with 5000 samples (central column) are compared with the test dataset (right column).

5.3 Latent-space assimilation of satellite observations

This section shows the results of the latent-space data assimilation for the Etna experiment (see Sect. 4.3). Figure 10a shows the ensemble mean of the model prior ($N_E = 128$), i.e. the ensemble of two-dimensional ash column mass at 18:00 UTC simulated with the dispersal model FALL3D. The corresponding SEVIRI retrievals at the same time instant are shown in Fig. 10b. The relatively complex spatial structure of the observed plume results from a pronounced clockwise rotation (veering) of its trajectory during the initial phase of the eruption. This change in flow direction is not captured by the driving meteorological dataset, which leads FALL3D to simulate ash transport under a largely uniform directional flow. Consequently, this case represents a challenging test for the data assimilation methodologies.

5.3.1 VAE prior

The simulated model prior was used to train a new VAE used to generate an augmented ensemble (VAE prior) by sampling from the latent space. The same procedure and neural network architecture used for the synthetic experiments was also employed for the Etna experiment, with a few exceptions: (i) in order to accommodate the increased size of the input data, the latent dimension was increased from $K = 16$ to $K = 32$, and a new convolutional layer was added to both the encoder and decoder; (ii) we found that the chequerboard pattern could be reduced by replacing transposed convolutional upsampling by a bilinear interpolation and, (iii) rather than sampling the latent variables from a standard normal distribution, the latent vectors were sampled from a multivariate normal distribution,

$$\mathbf{z}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{24}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ were estimated from the latent representations of the training ensemble. The VAE prior was then generated by sampling $N_E = 4096$ latent vectors $\mathbf{z}_i$ from this distribution and then decoded to obtain the VAE prior in the original ash column mass space.

The agreement with the satellite observations was evaluated using the metrics defined in Sect. 4.5. Table 4 compares the model prior and VAE prior against the satellite re-

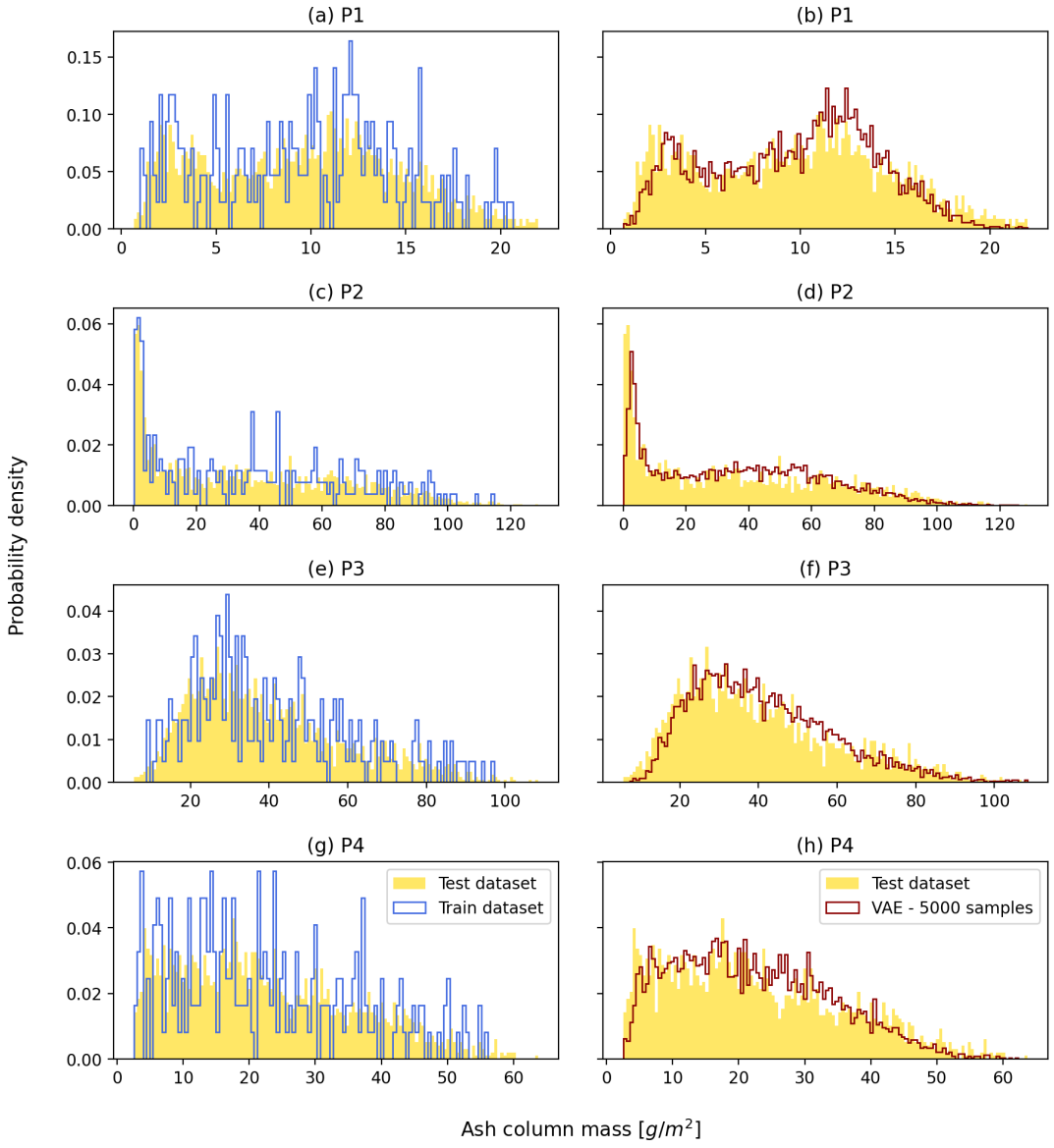


Figure 7. Histograms for the training dataset (256 samples, left column) and a VAE-generated ensemble (5000 samples, right column) at different locations. The results are compared against the test dataset (2048 samples), assumed to be the reference (yellow shaded area). For illustrative purposes, four sampling locations were selected: P1 at 15.5° W, 26.3° N **(a, b)**; P2 at 16.8° W, 25.9° N **(c, d)**; P3 at 15.0° W, 27.3° N **(e, f)**; and P4 at 15.5° W, 27.4° N **(g, h)**. See the Supplement for the point locations.

trievals. Metrics were computed for every ensemble member, and the ensemble average is reported in Table 4. Both model and VAE prior ensembles show very similar performance across all metrics, with the VAE prior yielding slightly better RMSE, RMSLE and Pearson correlation. These results indicate that the VAE-generated ensemble preserves the main characteristics of the model prior that are relevant to the comparison with observations. The results are also consistent with the previous experiment analysis, showing that the VAE can generate samples that are statistically comparable to those produced by the physics-based FALL3D model. These results provide a baseline for assessing the impact of DA in the next section, while also demonstrating that the VAE-generated ensemble provides a suitable prior for subsequent data assimilation experiments.

5.3.2 Data assimilation experiments

The analysis ensemble means obtained with the ETKF (Fig. 10c) and PF (Fig. 10d) show that both intensity and location of the maximum are corrected with respect to the prior. However, the analysis ensembles retain the lack of representation of the veering and wind shear present in the prior, resulting in an almost unidirectional and less extended plume.

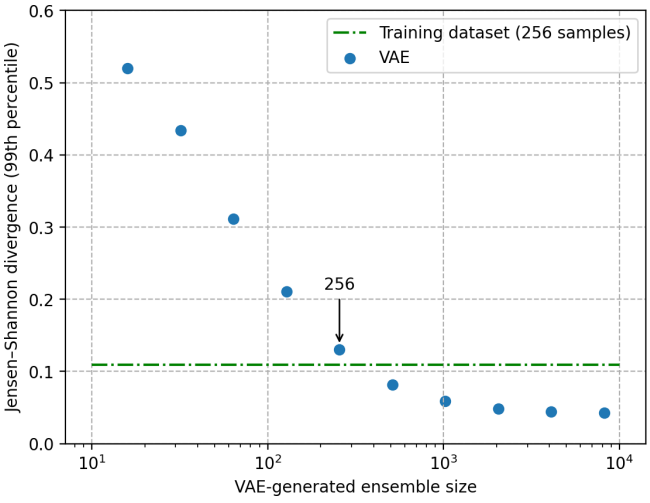


Figure 8. The Jensen-Shannon (JS) divergence was computed for each grid cell considering different sizes of the VAE-generated ensemble. The 99th percentile is reported as a worst-case measure of performance. Low values indicate a good approximation to the reference probability distribution. The result for the training dataset is also indicated for reference.

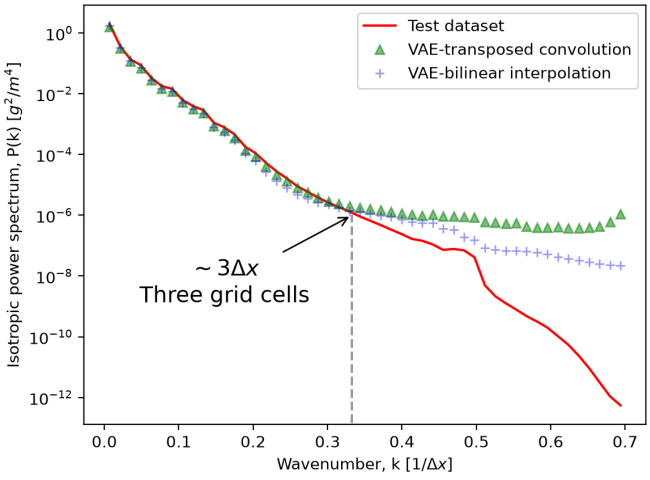


Figure 9. The isotropic power spectrum, i.e. radially averaged two-dimensional Power Spectral Density (PSD), for the test dataset and the VAE-generated samples obtained with transposed-convolution and bilinear-interpolation decoders. The generative model reproduces the large-scale spatial structure reasonably well, but exhibits excess high-wavenumber spectral power at scales smaller than approximately 3 Δx, indicating the introduction of artificial small-scale noise. This excess is more pronounced for the transposed-convolution decoder.

This illustrates an obvious but important limitation of the assimilation: the analysis can preferentially select and combine ensemble members that are more consistent with the observations, but cannot recover spatial structures that are not represented in the prior ensemble.

Table 4. Metrics used to quantify agreement with observations for the Etna data assimilation experiment.

	Prior (model)	Prior (VAE)	Posterior (ETKF)	Posterior (PF)
RMSE	180.317	173.562	145.696	126.903
RMSLE	1.561	1.491	1.228	1.210
Pearson correlation	0.315	0.339	0.700	0.667
POD	0.278	0.274	0.381	0.378
Jaccard index	0.229	0.227	0.325	0.333

The impact of the assimilation can be assessed in terms of the evaluation metrics reported in Table 4. These diagnostics are not intended as an independent validation, but rather to characterize the adjustment of the prior ensemble towards the observational constraints. The agreement with the observations improves for all metrics following assimilation. The RMSE is reduced by 16 % (ETKF) and 27 % (PF) relative to the VAE prior, while the RMSLE decreases by 18 % (ETKF) and 19 % (PF). The Pearson correlation approximately doubles for both methods. A larger relative improvement is obtained for the categorical metrics: the ETKF increases the POD and Jaccard index by 39 % and 43 %, respectively, while the corresponding increases for the PF are 38 % and 47 %. Overall, both assimilation approaches consistently adjust the ensemble towards the observations, with comparable performance across the different metrics considered.

The impact of the assimilation can also be assessed in terms of the ensemble spread. The ensemble spread quantifies the dispersion of the ensemble members around the mean and is computed as the ensemble standard deviation at each element, averaged over all elements. The resulting spreads for the prior and posterior ensembles are reported in Table 5. This metric is evaluated both in physical space and in latent space. The spread reduction in latent space is 10 % (ETKF) and 42 % (PF). The substantially lower PF spread results from particle-filter degeneracy, with nearly all posterior probability concentrated on only two particles. Kernel perturbation is subsequently used to increase ensemble diversity, with a regularization factor of $h = 0.2$ (see Sect. 3.3). Interestingly, the spread reduction in physical space is more pronounced for the ETKF (77 %) than for the PF (70 %), despite the substantially larger reduction observed for the PF in latent space.

The latent-space distributions of the prior and posterior ensembles are shown in Fig. 11 for selected pairs of latent dimensions, together with their corresponding histograms. In general, the prior distribution is well approximated by the standard normal distribution, as expected from the VAE construction. The ETKF produces a Gaussian-like posterior, whereas the PF analysis is clearly bimodal, reflecting the concentration of the posterior weights on two particles dur-

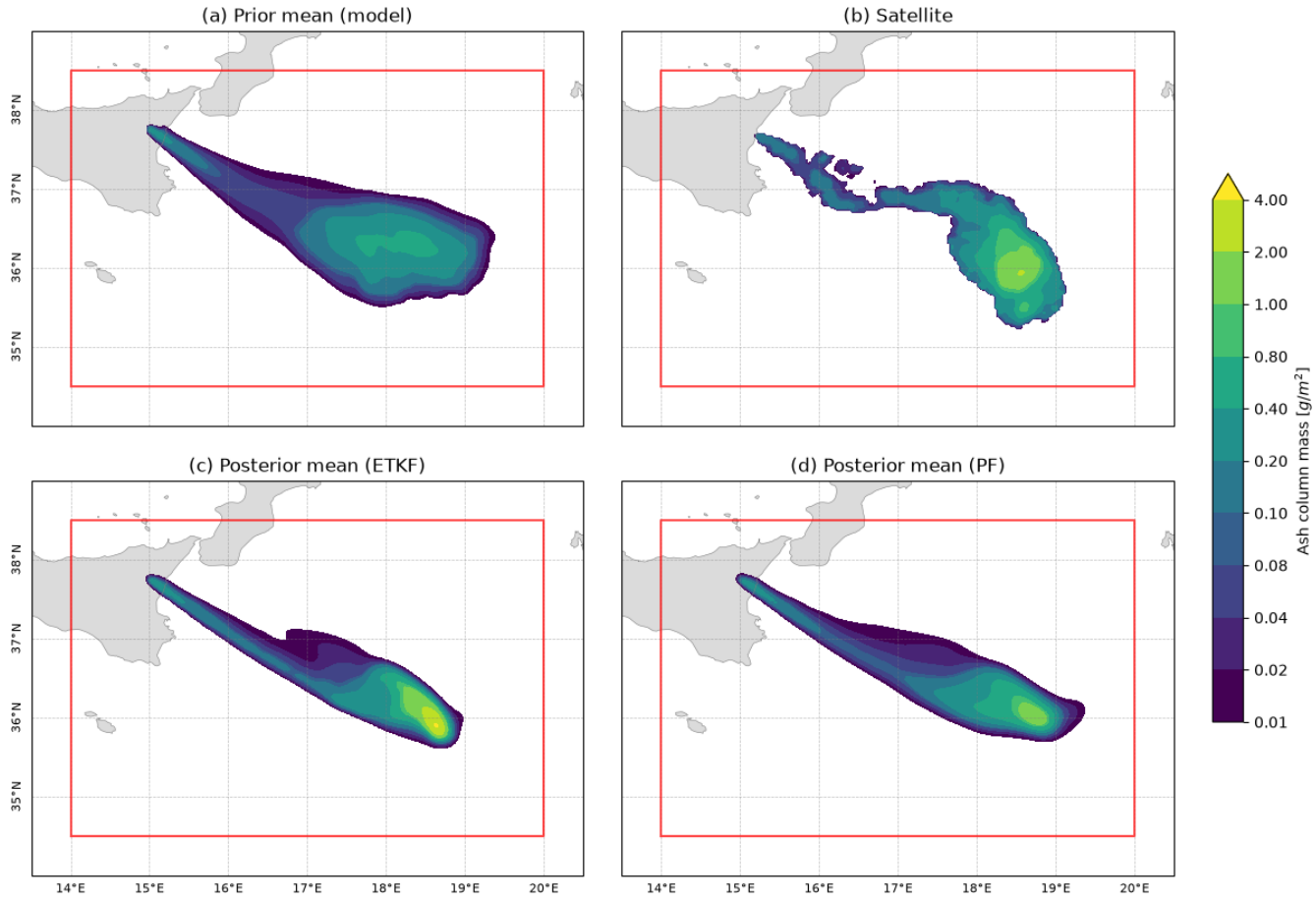


Figure 10. Spatial distribution of ash column mass at 18:00 UTC for the Etna 24 December 2018 case. **(a)** Ensemble mean of the model prior (128 members), **(b)** SEVIRI satellite retrieval, **(c)** ETKF posterior ensemble mean and, **(d)** PF posterior ensemble mean. The red box shows the extent of the FALL3D model domain.

Table 5. Ensemble spread of the prior and posterior ensembles in latent and physical space

	Latent-space spread	Physical-space spread
VAE Prior	0.999	33.974
ETKF Posterior	0.899	7.804
PF Posterior	0.582	10.351

ing the initial filter update. The subsequent kernel perturbation broadens these two modes, restoring some variability.

5.4 Discussion

In the context of this work, the VAE is viewed as a complement to the physics-based model. We showed that, making use of an ensemble of a few hundred simulations, the VAE can learn a compact, low-dimensional representation of the complex, high-dimensional variability captured by the physics-based simulations. This representation provides an alternative space in which the structure of the ensemble can be easily analysed and manipulated.

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.

Sampling in the latent space to generate an augmented ensemble is the most straightforward application of this methodology. Although it does not introduce new physical knowledge, the VAE can generate a much larger number of realizations at very low computational cost during inference, while retaining variability that is statistically consis-

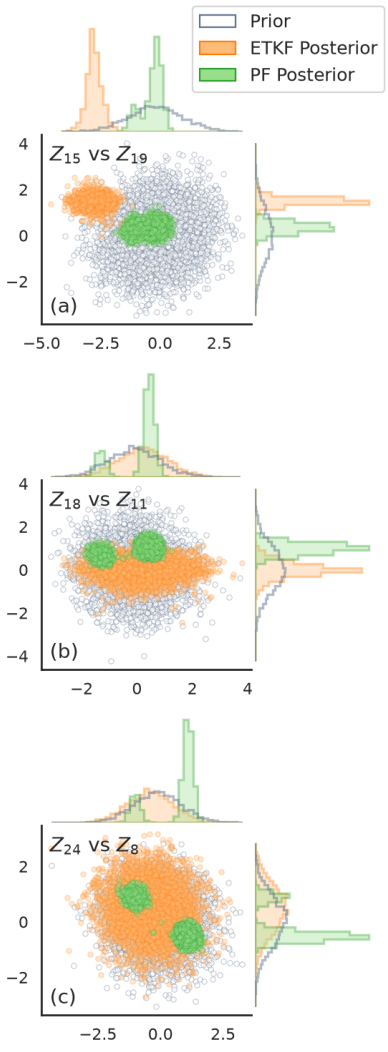


Figure 11. Latent-space distributions of the prior and posterior ensembles from the data assimilation experiments. Three selected pairs of the 32 latent dimensions are shown for illustration. The ensembles, each comprising 4096 members, are: VAE prior (white), ETKF posterior (orange), and PF posterior (green).

tent with that captured by the physics-based simulations. 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 2048 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. However, computational efficiency is arguably not the main advantage of the proposed approach.

The methodology presented in this work also offers potential advantages beyond ensemble augmentation. Increasing the ensemble size can reduce sampling errors and improve the estimation of statistical quantities that are particularly sensitive to small ensemble sizes, such as exceedance probabilities and other measures of uncertainty. Such estimates are particularly relevant for probabilistic hazard assessment and decision-making in crisis management. For example, as shown in Sect. 5.2.2, the VAE provides a better approximation to the reference distribution than the training ensemble in terms of the Jensen–Shannon (JS) divergence when the ensemble size exceeds the number of numerical simulations used for training.

Additionally, due to the generative nature of the VAE, instead of simply duplicating or resampling existing numerical simulations, as in a bootstrap approach, new realizations absent from the original dataset can be generated. This could potentially lead to a richer statistical representation of the variability captured by the physics-based model, particularly if some of the generated realizations correspond to regions of the distribution that are poorly represented in the training ensemble. This interpretation is consistent with the results presented in Sect. 3, where the VAE-generated ensemble shows better agreement with the observations than the model prior in terms of RMSE, RMSLE, and Pearson correlation (Table 4).

Finally and most importantly, the learned representation enables different operations to be performed in a substantially lower-dimensional space, including data assimilation, clustering, perturbation, conditional sampling, and other statistical techniques. These capabilities provide an excellent opportunity for further research. In this study, we have explored two latent-space data assimilation approaches that, in the formulation adopted here, would not be directly applicable in the original physical space. The assimilation experiments demonstrate the feasibility of performing data assimilation directly in the VAE latent space. Both the ETKF and PF adjust the augmented prior towards the satellite observations, improving the agreement across all considered evaluation metrics and reducing the ensemble spread without an excessive loss of ensemble variability. Although the two as-

similation approaches produce markedly different posterior distributions in latent space, their resulting posterior ensembles in physical space are relatively similar. The well-known particle-collapse problem of the PF was not fully resolved, as the posterior distribution remained concentrated around two particles, resulting in a bimodal distribution after the perturbation introduced by the kernel regularization. Applying such a naive perturbation directly in the physical space for a non-linear model would result in states that are inconsistent with the model dynamics. Instead, perturbing in latent space takes advantage of the learned mapping, allowing perturbed particles to be mapped back to physical space while retaining consistency with the physics-based simulations. This clearly illustrates one of the advantages of performing operations directly in the latent space.

Overall, the results presented in Sect. 5.3 demonstrate that latent-space data assimilation can effectively incorporate satellite observations to correct a large VAE-generated ensemble of two-dimensional volcanic ash column mass. 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. 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.

6 Conclusions

This study explored the potential of a hybrid strategy for volcanic-cloud modelling based on the integration of generative AI with physics-based simulations. A Variational Autoencoder (VAE) was trained on a relatively small ensemble of numerical simulations to learn a compact, low-dimensional representation of the high-dimensional problem. This methodology offers several potential advantages, including the generation of large ensembles at very low computational cost, the reduction of sampling errors and improved estimation of ensemble statistics, the generation of new samples beyond those contained in the training dataset, and the ability to perform statistical operations in a substantially lower-dimensional space, including latent-space data assimilation.

Two numerical experiments have been presented. A first synthetic experiment assessed the consistency of the VAE-generated ensemble using a controlled case characterized by multiple qualitatively different solutions, testing the ability of the VAE to capture this variability. 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. The results also highlight limitations in the representation of smaller-scale spatial structures. However, the reconstruction error is smaller than the errors introduced by the numerical model itself. In the second experiment, the 24 December 2018 Etna eruption was modelled and the VAE-augmented ensemble was compared with satellite retrievals. The agreement with observations was found very similar for the numerical and VAE-augmented ensembles, with the latter yielding better RMSE, RMSLE, and Pearson correlation coefficients, suggesting that the generation of new samples can lead to a better statistical representation of forecast uncertainty. The Etna experiment was also used to provide a proof-of-concept demonstration of latent-space data assimilation methodologies. A two-dimensional estimation problem was solved by assimilating satellite retrievals in the VAE latent space. The analysis ensemble showed improved agreement with the observations across all evaluated metrics, while significantly reducing the ensemble spread and retaining the ability to represent the spatial variability of the volcanic cloud and the physical consistency with the simulated prior.

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.

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/ (last access: 9 November 2026) and https://doi.org/10.5281/zenodo.17737081 (Folch et al., 2020; Prata et al., 2021). The code and data used for the synthetic and Etna experiments are available in the supplementary material repositories (https://doi.org/10.5281/zenodo.22759580, Mingari, 2026b), (https://doi.org/10.5281/zenodo.22759573, Mingari, 2026a).

Data availability. The data used in the synthetic and Etna experiments, including the corresponding model outputs and observational data, are publicly available in two Zenodo repositories: https://doi.org/10.5281/zenodo.22759580 (Mingari, 2026b) and https://doi.org/10.5281/zenodo.22759573 (Mingari, 2026a). The meteorological data used to drive the simulations were obtained

from the Copernicus European Regional ReAnalysis (CERRA), provided by the Copernicus Climate Change Service (C3S) Climate Data Store (CDS) (last access: 9 November 2026).

Supplement. The supplement related to this article is available online at https://doi.org/10.5194/acp-26-15953-2026-supplement.

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 contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. The authors bear the ultimate responsibility for providing appropriate place names. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

Acknowledgements. 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). We gratefully acknowledge Stefano Corradi for supplying the satellite retrieval data of the Etna eruption used in this work. Finally, we also thank the two anonymous reviewers for their valuable feedback and suggestions.

Financial support. 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.

The article processing charges for this open-access publication were covered by the CSIC Open Access Publication Support Initiative through its Unit of Information Resources for Research (URICI).

Review statement. This paper was edited by Peer Nowack and reviewed by two anonymous referees.

References

Arulampalam, M., Maskell, S., Gordon, N., and Clapp, T.: A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking, IEEE Trans. Signal Process., 50, 174–188, https://doi.org/10.1109/78.978374, 2002.

Bauer, P., Thorpe, A., and Brunet, G.: The quiet revolution of numerical weather prediction, Nature, 525, 47–55, https://doi.org/10.1038/nature14956, 2015.

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, https://doi.org/10.1029/2024JC022206, 2025.

Corradini, S., Guerrieri, L., Stelitano, D., Salerno, G., Scollo, S., Merucci, L., Prestifilippo, M., Musacchio, M., Silvestri, M., Lombardo, V., and Caltabiano, T.: Near Real-Time Monitoring of the Christmas 2018 Etna Eruption Using SEVIRI and Products Validation, Remote Sens., 12, https://doi.org/10.3390/rs12081336, 2020.

Corradini, S., Guerrieri, L., Brenot, H., Clarisse, L., Merucci, L., Pardini, F., Prata, A. J., Realmuto, V. J., Stelitano, D., and Theys, N.: Tropospheric Volcanic SO_2 Mass and Flux Retrievals from Satellite. The Etna December 2018 Eruption, Remote Sens., 13, https://doi.org/10.3390/rs13112225, 2021.

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, 114463, 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.

Doucet, A., de Freitas, N., and Gordon, N.: An Introduction to Sequential Monte Carlo Methods, in: Sequential Monte Carlo Methods in Practice, edited by: Doucet, A., de Freitas, N., and Gordon, N., 3–14, Springer, New York, https://doi.org/10.1007/978-1-4757-3437-9_1, 2001.

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 (last access: 9 November 2026), 2025.

Folch, A. and Mingari, L.: Data Assimilation of Volcanic Clouds: Recent Advances and Implications on Operational Forecasts, in: Applications of Data Assimilation and Inverse Problems in the Earth Sciences, edited by: Ismail-Zadeh, A., Castelli, F., Jones, D., and Sanchez, S., Special Publications of the International Union of Geodesy and Geophysics, 144–156, Cambridge University Press, Cambridge, https://doi.org/10.1017/9781009180412.010, 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, 741841, 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., Boehm, C., Brizuela, B., Brogi, F., Cabrera, E., Casarotti, E., Castro, M. J., Cerminara, M., Cirella, A., Cheptsov, A., Conejero, J., Costa, A., de la Asunción, M., de la Puente, J., Djuric, M., Dorozhinskii, R., Espinosa, G., Esposti-Ongaro, T., Farnós, J., Favretto-Cristini, N., Fichtner, A., Fournier, A., Gabriel, A.-A., Gallard, J.-M., Gibbons, S. J., Glimsdal, S., González-Vida, J. M., Gracia, J., Gregorio, R., Gutierrez, N., Halldorsson, B., Hamitou, O., Houzeaux, G., Jaure, S., Kessar, M., Krenz, L., Krischer, L., Laforet, S., Lanucara, P., Li, B., Lorenzino, M. C., Lorito, S., Løvholt, F., Macedonio, G., Macías, J., Marín, G., Martínez Montesinos, B., Mingari, L., Moguilny, G., Montellier, V., Monterrubio-Velasco, M., Moulard, G. E., Nagaso, M., Nazaria, M., Niethammer, C., Pardini, F., Pienkowska, M., Pizzimenti, L., Poiata, N., Rannabauer, L., Rojas, O., Rodriguez, J. E., Romano, F., Rudyy, O., Ruggiero, V., Samfass, P., Sánchez-Linares, C., Sanchez, S., Sandri, L., Scala, A., Schaeffer, N., Schuchart, J., Selva, J., Sergeant, A., Stallone, A., Taroni, M., Thrastarson, S., Titos, M., Tonelllo, N., Tonini, R., Ulrich, T., Vilotte, J.-P., Vöge, M., Volpe, M., Aniko Wirp, S., and Wössner, U.: The EU Center of Excellence for Exascale in Solid Earth (ChEESE): Implementation, results, and roadmap for the second phase, Future Gener. Comp. Sy., 146, 47–61, https://doi.org/10.1016/j.future.2023.04.006, 2023.

Folch, A., Costa, A., Macedonio, G., Mingari, L., and Pascual Saldaña, H.: FALL3D (Version 9.1.0), Zenodo [code], 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, Atmos. Chem. Phys., 16, 9189–9200, https://doi.org/10.5194/acp-16-9189-2016, 2016.

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, Atmos. Chem. Phys., 17, 1187–1205, https://doi.org/10.5194/acp-17-1187-2017, 2017.

Goodfellow, I., Bengio, Y., and Courville, A.: Deep Learning, MIT Press, ISBN 9780262035613, http://www.deeplearningbook.org (last access: 9 November 2026), 2016.

Gordon, N., Salmond, D., and Smith, A.: Novel approach to nonlinear/non-Gaussian Bayesian state estimation, IEE Proc.-F, Radar Signal Process., 140, 107–113, https://doi.org/10.1049/ip-f-2.1993.0015, 1993.

Hunt, B. R., Kostelich, E. J., and Szunyogh, I.: Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter, Physica D, 230, 112–126, https://doi.org/10.1016/j.physd.2006.11.008, 2007.

Husak, G. J., Michaelsen, J., and Funk, C.: Use of the gamma distribution to represent monthly rainfall in Africa for drought monitoring applications, Int. J. Climatol., 27, 935–944, https://doi.org/10.1002/joc.1441, 2007.

Kalnay, E.: Atmospheric modeling, data assimilation and predictability, Cambridge university press, https://doi.org/10.1017/CBO9780511802270, 2002.

Kingma, D. P. and Welling, M.: Auto-Encoding Variational Bayes, arXiv [preprint], https://doi.org/10.48550/arXiv.1312.6114, 2013.

Kitagawa, G.: Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models, J. Comput. Graph. Stat., 5, 1–25, https://doi.org/10.1080/10618600.1996.10474692, 1996.

Kliewer, A. J., Fletcher, S. J., Jones, A. S., and Forsythe, J. M.: Comparison of Gaussian, logarithmic transform and mixed Gaussian–log-normal distribution based 1DVAR microwave temperature–water-vapour mixing ratio retrievals, Q. J. Roy. Meteor. Soc., 142, 274–286, https://doi.org/10.1002/qj.2651, 2016.

Leutbecher, M. and Palmer, T.: Ensemble forecasting, J. Comput. Phys., 227, 3515–3539, https://doi.org/10.1016/j.jcp.2007.02.014, 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, 2014.

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

McInnes, L., Healy, J., and Melville, J.: UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, arXiv [preprint], https://doi.org/10.48550/arXiv.1802.03426, 2020.

Mingari, L.: Supplementary material for egusphere-2025-6097: Etna experiment, Zenodo [data set and code], https://doi.org/10.5281/zenodo.22759573, 2026a.

Mingari, L.: Supplementary material for egusphere-2025-6097: Synthetic experiment, Zenodo [data set and code], https://doi.org/10.5281/zenodo.22759580, 2026b.

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

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.

O'Neill, N. T., Ignatov, A., Holben, B. N., and Eck, T. F.: The lognormal distribution as a reference for reporting aerosol optical depth statistics; Empirical tests using multi-year, multi-site AERONET Sunphotometer data, Geophys. Res. Lett., 27, 3333–3336, https://doi.org/10.1029/2000GL011581, 2000.

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., and de' Michieli Vitturi, M.: Ensemble-Based Data Assimilation of Volcanic Ash Clouds from Satellite Observations: Application to the 24 December 2018 Mt. Etna Explosive Eruption, Atmosphere, 11, https://doi.org/10.3390/atmos11040359, 2020.

Pasmans, I., Chen, Y., Sebastian Finn, T., Bocquet, M., and Carrassi, A.: Ensemble Kalman filter in latent space using a variational autoencoder pair, Q. J. Roy. Meteor. Soc., 152, https://doi.org/10.1002/qj.70070, 2026.

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, arXiv [preprint], https://doi.org/10.48550/arXiv.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.

Pham, D. T.: Stochastic Methods for Sequential Data Assimilation in Strongly Nonlinear Systems, Mon. Weather Rev., 129, 1194–1207, https://doi.org/10.1175/1520-0493(2001)129<1194:SMFSDA>2.0.CO;2, 2001.

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, 105136, 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, https://doi.org/10.1038/s41598-018-24651-1, 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.

Ridal, M., Bazile, E., Le Moigne, P., Randriamampianina, R., Schimanke, S., Andrae, U., Berggren, L., Brousseau, P., Dahlgren, P., Edvinsson, L., El-Said, A., Glinton, M., Hagelin, S., Hopsch, S., Isaksson, L., Medeiros, P., Olsson, E., Unden, P., and Wang, Z. Q.: CERRA, the Copernicus European Regional Reanalysis system, Q. J. Roy. Meteor. Soc., 150, 3385–3411, https://doi.org/10.1002/qj.4764, 2024.

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, in: Modern Volcano Monitoring, edited by: Spica, Z. and Caudron, C., 365–397, Springer Nature Switzerland, Cham, https://doi.org/10.1007/978-3-031-86841-2_12, 2025.

Schimanke, S., Ridal, M., Le Moigne, P., Berggren, L., Undén, P., Randriamampianina, R., Andrae, U., Bazile, E., Bertelsen, A., Brousseau, P., Dahlgren, P., Edvinsson, L., El Said, A., Glinton, M., Hopsch, S., Isaksson, L., Mladek, R., Olsson, E., Verrelle, A., and Wang, Z. Q.: CERRA sub-daily regional reanalysis data for Europe on single levels from 1984 to present, Copernicus Climate Change Service (C3S) Climate Data Store (CDS) [data set], https://doi.org/10.24381/cds.622a565a, 2021.

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/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, 108046, 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.

van Leeuwen, P. J.: Particle Filtering in Geophysical Systems, Mon. Weather Rev., 137, 4089–4114, https://doi.org/10.1175/2009MWR2835.1, 2009.

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.

Wu, H. and Levinson, D.: The ensemble approach to forecasting: A review and synthesis, Transp. Res. Part C-Emerg. Technol., 132, 103357, https://doi.org/10.1016/j.trc.2021.103357, 2021.

Zidikheri, M. J. and Lucas, C.: A Computationally Efficient Ensemble Filtering Scheme for Quantitative Volcanic Ash Forecasts, J. Geophys. Res.-Atmos., 126, 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.