JGR Machine Learning and Computation

RESEARCH ARTICLE
10.1029/2025JH000819

Learning Low-Dimensional Representations of Ensemble Forecast Fields Using Autoencoder-Based Methods

Jieyu Chen[1,2], **Kevin Höhlein**[3], **and Sebastian Lerch**[1,4,5]

[1]Institute of Statistics, Karlsruhe Institute of Technology, Karlsruhe, Germany, [2]School of Mathematics and Statistics, Nanjing University of Information Science and Technology, Nanjing, China, [3]TUM School of Computation, Information and Technology, Technical University of Munich, Munich, Germany, [4]Department of Mathematics and Computer Science, Marburg University, Marburg, Germany, [5]Heidelberg Institute for Theoretical Studies, Heidelberg, Germany

Key Points:

- We introduce novel dimensionality reduction approaches specifically tailored to ensembles of gridded simulation fields
- Our method translates raw ensemble forecast fields into a compressed distributional latent representation
- The distributional representations summarize key information about the spatial structure and the variability encoded in the ensemble

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

J. Chen,
jieyu.chen@nuist.edu.cn

Citation:

Chen, J., Höhlein, K., & Lerch, S. (2025). Learning low-dimensional representations of ensemble forecast fields using autoencoder-based methods. *Journal of Geophysical Research: Machine Learning and Computation*, *2*, e2025JH000819. https://doi.org/10.1029/2025JH000819

Received 16 JUN 2025
Accepted 15 OCT 2025

Author Contributions:
Conceptualization: Sebastian Lerch
Funding acquisition: Sebastian Lerch
Methodology: Jieyu Chen, Kevin Höhlein, Sebastian Lerch
Project administration: Sebastian Lerch
Resources: Sebastian Lerch
Software: Jieyu Chen, Kevin Höhlein
Supervision: Sebastian Lerch
Validation: Jieyu Chen
Visualization: Jieyu Chen
Writing – original draft: Jieyu Chen, Sebastian Lerch
Writing – review & editing: Jieyu Chen, Kevin Höhlein, Sebastian Lerch

Abstract

Large-scale numerical simulations often produce high-dimensional gridded data, which is challenging to process for downstream applications. A prime example is numerical weather prediction, where atmospheric processes are modeled using discrete gridded representations of the physical variables and dynamics. Uncertainties are assessed by running the simulations multiple times, yielding ensembles of simulated fields as a high-dimensional stochastic representation of the forecast distribution. The high dimensionality and large volume of ensemble data sets imposes major computing challenges for subsequent forecasting stages. Data-driven dimensionality reduction techniques could help to reduce the data volume before further processing by learning meaningful and compact representations. However, existing dimensionality reduction methods are typically designed for deterministic and single-valued inputs, and thus they cannot handle ensemble data from multiple randomized simulations. In this study, we propose novel dimensionality reduction approaches specifically tailored to the format of ensemble forecast fields. We present two alternative frameworks, which yield low-dimensional representations of ensemble forecasts while respecting their probabilistic character. The first approach derives a distribution-based representation of an input ensemble by applying standard dimensionality reduction techniques in a member-by-member fashion and merging the member representations into a joint parametric distribution model. The second approach achieves a similar representation by encoding all members jointly using a tailored variational autoencoder. We evaluate and compare both approaches in a case study using 10 years of temperature and wind speed forecasts over Europe. The approaches preserve key spatial and statistical characteristics of the ensemble and enable efficient generation of additional member forecast fields.

Plain Language Summary

Modern weather forecasts rely on large-scale physical simulations that model the atmosphere over gridded spatial fields. These simulations are often run multiple times with slightly different configurations to account for forecast uncertainty, producing the ensemble forecast. While this approach improves reliability, it also generates massive volumes of high-dimensional data that are difficult to be stored, processed, and used in downstream applications. In this study, we developed novel machine learning methods to reduce the size of these gridded ensemble forecast data sets without losing crucial uncertainty information. Our methods are specifically designed to handle the peculiar characteristics of ensemble forecasts, whereas traditional data compression techniques cannot be directly applied to this type of data. We proposed two types of approaches: one that compresses each simulation individually and then combines them and another that compresses all simulations together using a variational autoencoder-based neural network. We applied these methods to 10 years of temperature and wind speed forecasts across Europe. The results show that our approaches preserve key spatial patterns and uncertainty characteristics, enabling more efficient use of ensemble forecasts while maintaining their essential information.

1. Introduction

Large-scale physics-based models are used across environmental sciences for prediction and modeling. A particularly important example is numerical weather prediction (NWP) models, where atmospheric processes are represented via partial differential equations. The forecast quality of NWP models has improved tremendously in recent decades due to continued scientific and technological advances (Bauer et al., 2015). Nowadays, NWP models are often run in an ensemble mode to quantify forecast uncertainty. Thereby, a collection of predictions of future weather states is obtained by running the model several times with varying initial conditions and perturbed model physics. Conceptually, these ensemble members are considered equally probable realizations of an unknown probability distribution. In many current NWP systems, the ensemble members are therefore considered to be interchangeable.

Due to their continuously increasing spatial and temporal resolution, ensemble weather forecasting models produce large amounts of data. However, it is challenging to process such high-dimensional and complex data in applications relying on weather predictions as inputs. Examples include weather forecasting applications such as postprocessing and analog forecasting and downstream applications such as hydrological and energy forecasting models. Therefore, summarizing relevant information from meteorological input data across space and time via learning low-dimensional representations is of interest beyond just reducing the amount of data that needs to be stored.

One example is ensemble postprocessing, which aims at correcting systematic errors of NWP ensemble predictions via statistical or machine learning (ML) models (Vannitsem et al., 2021). Postprocessing models use ensemble predictions of relevant meteorological variables as inputs and produce corrected probabilistic forecasts in the form of probability distributions as their output. While recent ML-based approaches have enabled the incorporation of many predictor variables (Chen et al., 2024; Rasp & Lerch, 2018; Schulz & Lerch, 2022) and there exist first spatial postprocessing approaches (Chapman et al., 2022; Grönquist et al., 2021; Horat & Lerch, 2024; Scheuerer et al., 2020; Veldkamp et al., 2021), most postprocessing models still tend to operate on localized predictions at individual stations or grid point locations. However, the restriction to localized predictions prevents the incorporation of predictability information from large-scale spatial structures, including weather regimes. While such structures are inherently represented in physically consistent forecast fields from NWP models and have been demonstrated to provide relevant information on conditional, flow-dependent error characteristics of NWP forecasts (Allen et al., 2021; Rodwell et al., 2018), directly utilizing them as inputs to postprocessing models along with station-based variables is challenging due to their high dimensionality. To address this limitation, Lerch and Polsterer (2022) propose the use of convolutional autoencoders (AEs) to learn low-dimensional latent representations of the high-dimensional gridded forecast fields and demonstrate that using the encoded representations as additional predictors to augment an NN-based postprocessing model with information about the spatial structure of relevant forecast fields helps to improve predictive performance. However, Lerch and Polsterer (2022) only utilize encoded representation of the mean ensemble field, where all ensemble member forecasts are averaged at every grid point. One potential drawback is that the mean field will be notably smoother than forecast fields from individual members. More importantly, however, such approaches ignore the underlying probabilistic information available in the ensemble simulations, which can be seen as samples from a multivariate probability distribution.

Our overarching aim is to propose dimensionality reduction methods to learn low-dimensional representations of ensemble forecast fields, which respect the inherently probabilistic nature of the input data. A variety of dimensionality reduction methods are available, ranging from classical principal component analysis (PCA; Jolliffe & Cadima, 2016; Pearson, 1901) to neural network (NN)-based AE methods (Bourlard & Kamp, 1988; Hinton & Salakhutdinov, 2006; Hinton & Zemel, 1993; Kramer, 1991). However, the application of existing dimensionality reduction methods to ensemble forecast fields is not straightforward, since they tend to be tailored to deterministic input data. To the best of our knowledge, the problem of learning representations of ensemble simulation data has not been considered thus far, potentially since this type of data is somewhat specific to environmental modeling. The key design considerations, which also constitute the main challenges of this study, lie in two aspects. First, the representations must capture both the large-scale spatial structures of the forecast fields and the forecast uncertainty, in the form of variability across ensemble members, within a probabilistic framework, whereas existing approaches typically yield only deterministic representations. Second, the encoded probabilistic representations should allow for the generation of synthetic ensemble member fields that are, in principle, indistinguishable from randomly selected members of the original ensemble, which can be achieved by decoding samples drawn from the latent distribution. Therefore, we aim to develop dimensionality reduction approaches that learn distributional representations in the latent space for an ensemble of forecast fields to address these challenges.

To achieve this, we propose two approaches, one based on existing dimensionality reduction methods and one utilizing variational autoencoder (VAE; Kingma, 2013) architectures. The former is an extension of existing dimensionality reduction models with deterministic latent code and can be summarized as a two-step framework.

In the first step, a dimensionality reduction model (e.g., PCA or an AE model) is employed to learn low-dimensional representations for each member of the ensemble forecast fields. In the second step, a multivariate Gaussian distribution in the latent space is fitted to the encoded representations of all ensemble members. This distribution serves as a compact probabilistic representation of the entire ensemble and can be used to reconstruct ensemble members that are statistically indistinguishable from the original members. This is achieved by reverting the encoding process, that is, drawing independent samples from the fitted distribution and applying the reverse step of the dimensionality reduction model (e.g., inverse PCA transform or the decoder of AE). The VAE-based framework, which we propose as a conceptually distinct alternative, utilizes a tailored VAE model that jointly considers all ensemble members as one input and provides a distributional ensemble representation as the encoder posterior distribution defined on the VAE's latent space. A key design consideration is that the proposed VAE model should respect the interpretation of ensemble members as interchangeable samples from an unknown, multivariate probability distribution. Notably, the obtained distributional representation should be independent of any (arbitrary) ordering in which the ensemble members are sampled, held in memory, and supplied to the VAE. To this end, we use an invariant VAE (iVAE) architecture designed to be invariant to the reordering of ensemble members. In contrast to the PCA and AE-based approaches, the iVAE model thus does not encode deterministic latent representations but directly learns a probability distribution in the latent space while treating ensemble members as invariant inputs.

We systematically compare the two approaches in two case studies on ensemble forecast fields covering a region that roughly corresponds to Europe. We focus on 2-day ahead forecasts of temperature and wind speed, utilizing 10 years of daily forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF). To that end, we discuss appropriate evaluation approaches for the problem at hand and consider an exemplary analysis of the encoded representations.

The remainder of the paper is structured as follows. Section 2 provides an overview of the data set, and Section 3 introduces the proposed two-step and iVAE approaches to learn distributional representations of ensemble forecast fields. The evaluation methods and main results are presented in Section 4, followed by conclusions and discussions in Section 5. Python code with implementations of all approaches is available online (https://github.com/jieyu97/invariantVAE).

2. Data

We focus on daily ensemble forecasts from the ECMWF's 50-member ensemble on a spatial domain roughly covering the European continent (−10°–30° E and 30°–70° N). The forecasts are available as gridded fields with regular $0.5° \times 0.5°$ resolution in latitude and longitude. This results in $81 \times 81 (= 6561)$ grid points over Europe. The forecasts are initialized daily at 00 UTC with a forecast lead time of 48 hr. We retrieve forecast data for all days in the time period from 3 January 2007 to 2 January 2017 and split the data into nonoverlapping parts for training (3 January 2007–31 December 2014), validation (1 January 2015–31 December 2015), and testing (remainder).

For brevity, we select four exemplary meteorological variables as the basis of our evaluation: temperature at 2 m (`t2m`) in Kelvin, zonal wind at 10 m (`u10`) in meter per second, meridional wind at 10 m (`v10`), and geopotential height at 500 hPa (`z500`). Given the overall similarity of results across the four variables and the better representativeness of `t2m` and `u10` for weather forecasting, we focus on these two variables in the presentation of results.

For each weather variable, we apply standard normalization to the raw ensemble forecast data for more stable training of NN models. The data are standardized by subtracting a global mean and dividing by a global standard deviation, both of which are computed over the entire data set. The parameters are computed separately for each weather variable using the data from all grid points in the domain and all samples in the training data set.

3. Learning Distributional Representations of Ensemble Forecast Fields

This section first introduces required mathematical notation and outlines the problem to be addressed and then presents two different frameworks for learning distributional representations of ensembles of spatial fields.

3.1. Mathematical Notations and Problem Formulation

Throughout this paper, we aim to find lower-dimensional probabilistic representations for ensembles of spatial forecast fields that capture both the structural and uncertainty information of the ensemble while reducing data complexity. The required representations are learned in a data-driven way using suitable statistical models for encoding and decoding the inputs. Due to the stochastic characteristic of ensembles, we focus on distributional representations, which express the ensemble information through a suitably parameterized probability distribution defined on a low-dimensional latent space. Such probabilistic representations offer two key conceptual advantages: first, they encode additional uncertainty information in the latent space, which can be valuable for downstream tasks that cannot directly handle high-dimensional data, and second, they allow for the computationally efficient generation of numerous ensemble member fields, which are theoretically interchangeable with the original ones, by decoding samples drawn from the latent distributions. The problem can thus be considered as a dimensionality reduction task with distribution-based embeddings that capture both the large-scale spatial structure of the forecast fields and the variability among ensemble members.

For a specific weather variable (e.g., 2-m temperature, `t2m`) and time t, we denote the 50-member ensemble forecast by $\boldsymbol{X}^{\texttt{t2m},t} = \{\boldsymbol{X}_m^{\texttt{t2m},t}\}_{m=1}^{50}$, wherein $\boldsymbol{X}_m^{\texttt{t2m},t} \in \mathbb{R}^{d_{\text{data}}}$ represents the m-th member forecast field. The superscripts `t2m` and t will typically be omitted for brevity. We then write $\boldsymbol{X} = \{\boldsymbol{X}_m\}_{m=1}^{50}$ to denote the ensemble forecast for a given variable and a given time, for example, to refer to a data point as one training example. The proposed dimensionality reduction methods process one meteorological variable at a time. Therefore, each forecast field $\boldsymbol{X}_m$ comprises scalar-valued forecast data for 81 × 81 grid locations, resulting in $d_{\text{data}} = 6561$. The 50 ensemble members are interpreted as independent samples from an unknown but identical multivariate probability distribution $\mathcal{P}$, which captures the uncertainty about the predicted weather state as follows:

$$\boldsymbol{X}_m \sim \mathcal{P} \quad \text{for } m \in \{1, \dots, 50\}.$$

Each dimensionality reduction consists of an encoding part $\mathtt{E}$, a decoding part $\mathtt{D}$, and a latent space $\mathbb{R}^{d_{\text{latent}}}$, which hosts the encoded representations. The encoding part learns to translate the input ensemble into a representative distribution $\mathcal{D}$ in the latent space, that is,

$$\mathcal{D} = p(\cdot|\theta = \mathbb{E}(\boldsymbol{X})),$$

and the decoding part is trained to reconstruct an ensemble of forecast fields $\tilde{\boldsymbol{X}} = \{\tilde{\boldsymbol{X}}_n\}_{n=1}^{N}$ based on an ensemble of samples $\boldsymbol{z} = \{\boldsymbol{z}_n\}_{n=1}^{N}$, drawn from $\mathcal{D}$, that is,

$$\tilde{\boldsymbol{X}} = \mathtt{D}(\boldsymbol{z}),$$

wherein $\boldsymbol{z}_n \sim \mathcal{D}$ for $n \in \{1, \dots, N\}$. While N—the size of the reconstructed ensemble—can be arbitrary, in general, we mainly consider the case $N = 50$, which matches the number of members in the original ensemble forecast. We note, however, that even in this case, $\tilde{\boldsymbol{X}}_n$ and $\boldsymbol{X}_m$ usually do not correspond, even if they have the same subscript value, since $\tilde{\boldsymbol{X}}_n$ is decoded from a random sample $\boldsymbol{z}_n$, which is not necessarily the adequate latent representation of $\boldsymbol{X}_m$ with $m = n$. The member-wise reconstruction of the original ensemble will be denoted as $\hat{\boldsymbol{X}} = \{\hat{\boldsymbol{X}}_m\}_{m=1}^{50}$.

Additionally, we usually have that the latent dimension $d_{\text{latent}} \ll d_{\text{data}}$. Therefore, d_{latent} controls the compression level of the methods, that is, how much information from each ensemble field remains in the latent representations. As a hyperparameter of the proposed methods, the latent dimension can be adapted to the needs of downstream tasks. In the presented case studies, we restrict our focus to lower latent dimensions ranging from 2 to 32. The information content of the representation is furthermore affected by the parametric form of the latent distribution $\mathcal{D}$, which is another design choice within the proposed method. We will assume $\mathcal{D}$ to be Gaussian, that is, $\mathcal{D} = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with parameters $\boldsymbol{\mu} \in \mathbb{R}^{d_{\text{latent}}}$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{d_{\text{latent}} \times d_{\text{latent}}}$ representing the distribution mean and the covariance matrix, respectively.

The ultimate goal of the proposed learning framework is finding suitable mappings $\mathtt{E}$ and $\mathtt{D}$ such that for multisamples z from $\mathcal{D}$, the reconstructed field ensemble $\mathtt{D}(z)$ becomes statistically indistinguishable from ensemble members sampled directly from the forecast distribution $\mathcal{P}$. The key challenge therein lies in performing dimensionality reduction on the space of probability distributions, which are represented through stochastically sampled ensembles of forecast fields. In this setting, the representations of each data point (i.e., ensemble $\boldsymbol{X}$) only convey incomplete and stochastic information about the underlying data (i.e., forecast distribution $\mathcal{P}$). This is in stark contrast to the assumption of standard dimensionality reduction problems, which presuppose complete and deterministic data representations. We propose two different approaches to address this problem, leveraging statistical and ML methods, which will be introduced in the following sections.

3.2. Two-Step Dimensionality Reduction Approaches

Our input data are 50-member ensembles of spatial forecast fields $\boldsymbol{X} = \{X_m\}_{m=1}^{50}$. The most straightforward approach is to treat all members collectively as one input $\boldsymbol{X}$ and utilize existing dimensionality reduction methods. However, treating the ensemble members jointly ignores their nature as interchangeable samples drawn from an identical distribution, leading to a deterministic latent representation for the entire ensemble. This conflicts with our goal of learning a representative low-dimensional distribution for the ensemble of forecast fields. Furthermore, the variabilities among different ensemble members are often less distinct than those among different grid locations in the spatial forecast fields. Consequently, the encoded deterministic representation primarily captures the spatial structure in the data. Initial experiments with existing techniques that learn a deterministic low-dimensional representation of all ensemble members jointly as one input indicated that the reconstructed forecast fields primarily approximate the ensemble mean and fail to reproduce any variability across individual ensemble members.

To address the probabilistic nature of the ensemble forecast fields and preserve uncertainty information, we propose a two-step framework to identify a latent distribution capturing both general spatial structure and variability within the ensemble. This framework builds on existing methods, which are used to reduce the dimension of each ensemble member separately before merging the per-member representations into a distributional form. Specifically, we assume that a given standard dimensionality reduction approach provides a mapping f, which maps a data item to its reduced representation, and a reconstruction function g, which restores a data item based on its latent code.

To encode an ensemble, we proceed by treating each member forecast field separately to obtain its deterministic low-dimensional representations as

$$\hat{z}_m = f(\boldsymbol{X}_m) \quad \text{for } m \in \{1, \dots, 50\}.$$

This yields an ensemble of latent representations to which we can fit a d_{latent}-dimensional Gaussian distribution in the latent space,

$$\mathcal{D} = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \text{with } \boldsymbol{\mu} = \frac{1}{50}\sum_{m=1}^{50} \hat{z}_m, \text{ and } \boldsymbol{\Sigma} = \operatorname{Var}(\hat{z}_1, \dots, \hat{z}_{50}).$$

Therein, $\boldsymbol{\mu}$ is the estimated mean vector and $\boldsymbol{\Sigma}$ is the estimated covariance matrix. This corresponds to the intended low-dimensional probabilistic representation.

Ensemble members are reconstructed by decoding samples drawn from $\mathcal{D}$, where the ensemble size N can be chosen arbitrarily, yielding

$$\tilde{X} = \{g(z_n)\}_{n=1}^{N}, \quad \text{with } z_n \sim \mathcal{D}, \quad \text{for } n \in \{1, \dots, N\}.$$

These newly generated forecast fields can be considered to follow the same distribution as the member-wise reconstructions of the input ensemble members, $\hat{\boldsymbol{X}}_m = g(\hat{z}_m)$, for $m \in \{1, \dots, 50\}$.

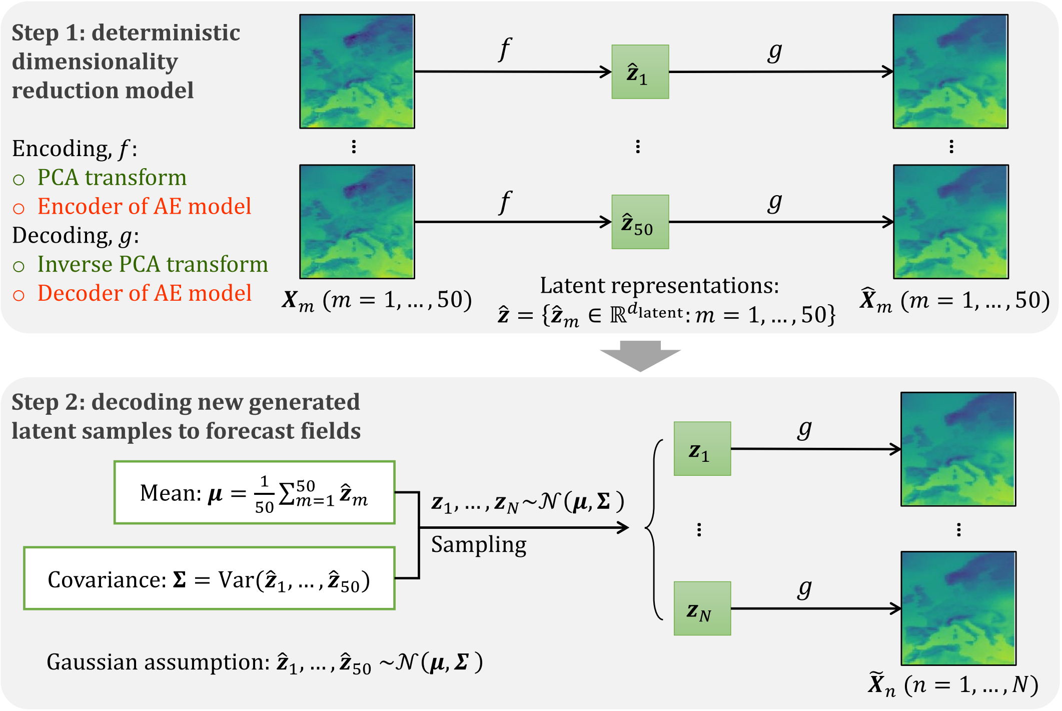
Step 1: deterministic dimensionality reduction model
Encoding, f:
○ PCA transform
○ Encoder of AE model
Decoding, g:
○ Inverse PCA transform
○ Decoder of AE model
f
g
$\hat{\boldsymbol{z}}_1$
$\hat{\boldsymbol{z}}_{50}$
$\boldsymbol{X}_m$ $(m = 1, \ldots, 50)$
Latent representations:
$\hat{\boldsymbol{z}} = \{\hat{\boldsymbol{z}}_m \in \mathbb{R}^{d_{\text{latent}}}: m = 1, \ldots, 50\}$
$\hat{\boldsymbol{X}}_m$ $(m = 1, \ldots, 50)$
Step 2: decoding new generated latent samples to forecast fields
Mean: $\boldsymbol{\mu} = \frac{1}{50}\sum_{m=1}^{50} \hat{\boldsymbol{z}}_m$
Covariance: $\boldsymbol{\Sigma} = \mathrm{Var}(\hat{\boldsymbol{z}}_1, \ldots, \hat{\boldsymbol{z}}_{50})$
$\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
Sampling
$\boldsymbol{z}_1$
$\boldsymbol{z}_N$
g
g
Gaussian assumption: $\hat{\boldsymbol{z}}_1, \ldots, \hat{\boldsymbol{z}}_{50} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
$\tilde{\boldsymbol{X}}_n$ $(n = 1, \ldots, N)$

Figure 1. Schematic overview of the two-step dimensionality reduction methods based on principal component analysis and autoencoder models.

We focus on two practical implementations of this approach using PCA and NN-based AE as the underlying algorithms. A schematic illustration of the AE approach is provided in Figure 1. PCA and AE are introduced in the following sections.

3.2.1. Principal Component Analysis Approach

Principal component analysis (PCA) is a linear method that finds the projections of data onto the principal components, which capture the largest variation in the data. The basic idea is to project all samples into a new coordinate system, where the axes (principal components) are determined by the direction along which projections have the largest variance in descending order. For a data set of d_{data} dimensions, the number of principal components is also d_{data}, and the dimensionality reduction is conducted by taking the projections onto only the first d_{latent} principal components. The reversibility of the PCA transform allows data to be reconstructed from the low-dimensional representations. To provide a benchmark for comparison, we employ PCA in the two-step framework and refer to it as the PCA-based approach.

The parameters of the PCA transformation are estimated from the training data, with each of the 50 original ensemble members included as a separate training instance. The resulting transformation is then applied directly to the test data, ensuring consistency of the principal components across ensemble members and enabling their coherent aggregation for further analysis. A potential concern with this joint estimation approach is that the leading principal components might reflect different sources of variability, such as spatial or member-wise effects. In our analysis, however, we find that the dominant modes of variation correspond to spatial patterns, while variability across ensemble members is comparatively minor. Therefore, the PCA-based representations tend to capture systematic features of the atmospheric state shared across ensemble members, rather than member-specific noise, which aligns with the goal of our two-step framework.

In many real-world data sets, linear transformations are not adequate to compress key information, and a variety of nonlinear dimensionality reduction techniques have been proposed. Examples include kernel PCA (Schölkopf et al., 1997), Isomap (Tenenbaum et al., 2000), and locally linear embedding (LLE; Roweis and Saul (2000)). While these methods provide effective low-dimensional representations, the process of converting them back to the original data space introduces additional challenges. The inverse transformations for those nonlinear techniques often require additional training procedures, as exemplified by the preimage problem for kernel PCA (Kwok & Tsang, 2004; Mika et al., 1998). AE models, on the other hand, have emerged as a more flexible and now widely used alternative in reconstructing data from latent features. Since PCA can be considered as a linear case of a simple NN, it is a natural reference method.

3.2.2. AE Neural Network Approach

AEs are NN models for unsupervised learning that aim to replicate the input as their output. A typical AE features an internal bottleneck layer with fewer nodes than the input and output layers, dividing the network into two distinct components, the encoder and the decoder. This bottleneck imposes a constraint that makes it more difficult for the network to simply memorize every detail of the input. The encoder f and the decoder g can be formulated as two mappings, following the notations in Section 3.1 as follows:

$$f(\boldsymbol{X}_m) = \hat{z}_m, \quad g(\hat{z}_m) = \hat{\boldsymbol{X}}_m, \quad \text{for } \boldsymbol{X}_m, \hat{\boldsymbol{X}}_m \in \mathbb{R}^{d_{\text{data}}},\ \hat{z}_m \in \mathbb{R}^{d_{\text{latent}}} \text{ and } m \in \{1, \dots, 50\}.$$

The encoder network f maps one input forecast field $\boldsymbol{X}_m$ to its latent representation $\hat{z}_m$ from the bottleneck layer, with d_{latent} typically much smaller than d_{data}. The decoder network g maps one latent representation $\hat{z}_m$ back to the corresponding reconstruction $\hat{\boldsymbol{X}}_m$, the output of the AE, which aims to reproduce the input $\boldsymbol{X}_m$. Training AEs involves minimizing differences between the deterministic input and output, often using mean square error (MSE) as a loss function.

The deterministic latent code $\hat{z}_m$ obtained from the encoder naturally functions as a compact representation of the input, capturing essential features needed for the decoder to reconstruct the original data. In addition to dimensionality reduction applications (Hinton & Salakhutdinov, 2006; W. Wang et al., 2014; Y. Wang et al., 2016), AE models have also found applications in other domains, such as anomaly detection (Sakurada & Yairi, 2014; Zhou & Paffenroth, 2017) and image denoising (Gondara, 2016). AEs exist in many variants, developed for different applications, including, for example, sparse AEs for classification tasks (Baccouche et al., 2012).

Our AE model for the AE-based dimensionality reduction approach is a shallow NN utilizing fully connected dense layers in both the encoder and the decoder. The model is trained to minimize the mean absolute error between the input forecast field and the reconstructed field obtained as output. Model parameters are optimized on the training and validation data including all ensemble members as separate instances, analogous to the setup used for PCA. Hyperparameter tuning is performed using the Bayesian optimization algorithm HyperBand (Li et al., 2018) implemented in the `Ray Tune` Python library (Liaw et al., 2018). The final AE model configuration consists of layers with sizes "6,561 – 4,096 - d_{latent}" in the encoder and "d_{latent} − 4,096 − 6,561" in the decoder, where 6,561 refers to the size of the input and output layers corresponding to the flattened grid, 4,096 represents the size of the hidden layers, and d_{latent} corresponds to the size of the bottleneck layers that captures the latent representations. The LeakyReLU activation function is applied in the hidden layers of both the encoder and the decoder. We utilize the AdamW optimizer (Loshchilov & Hutter, 2019) with a learning rate decay scheduler starting from 10^{-4} to stabilize the training process. Minibatch training is employed with a batch size of 1,024 to enhance training efficiency, and samples in all batches are randomly shuffled in each training epoch. To prevent overfitting, an early stopping criterion with a patience of 20 epochs on the validation loss is applied. We also investigated more sophisticated frameworks for the encoder and the decoder during initial experiments, including convolutional layers with residual blocks (He et al., 2016), and a vision transformer (ViT)-based (Dosovitskiy et al., 2021) architecture. However, these more complex approaches did not yield notable improvements, and we prioritize a conceptually simpler framework with only dense layers for our NN models but note that future improvements might be possible with alternative architectures.

3.3. Invariant Variational Autoencoder Approach

The two-step framework developed for ensemble forecast fields can, in principle, be generalized to other dimensionality reduction techniques beyond PCA and AE. However, a conceptual disadvantage of the approach is the assumption of Gaussian-distributed representations in the latent space, which is somewhat decoupled from the training process.

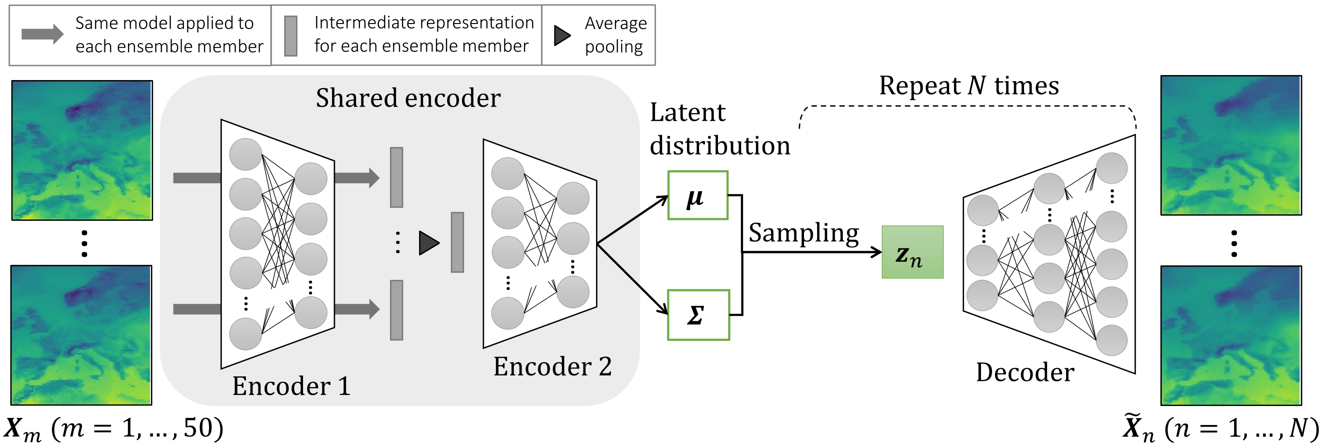
Same model applied to each ensemble member
Intermediate representation for each ensemble member
Average pooling
Shared encoder
Encoder 1
Encoder 2
Latent distribution
$\boldsymbol{\mu}$
$\boldsymbol{\Sigma}$
Sampling
Repeat N times
$\boldsymbol{z}_n$
Decoder
$\boldsymbol{X}_m\ (m = 1, \ldots, 50)$
$\tilde{\boldsymbol{X}}_n\ (n = 1, \ldots, N)$

Figure 2. Schematic illustration of the invariant variational autoencoder model.

To address this limitation, we propose an ensemble-invariant framework based on VAEs. VAEs (Kingma, 2013) are generative ML methods that leverage variational inference to learn a probabilistic representation in latent space. In contrast to standard AEs, VAEs connect an encoder network to its decoder through a probabilistic latent space, which corresponds to the parameters of a prespecified probability distribution. Thereby, the encoder network maps input samples to parameters of the latent space distribution, and the decoder network maps samples drawn from the distribution in the latent space back to the data space by generating new data points decoded from the samples. The reparameterization trick (Kingma & Welling, 2019) enables the simultaneous training of the encoder and the decoder using backpropagation by transforming the sampling process to make it differentiable. VAEs have been widely applied in various domains, including image generation, denoising, and inpainting (An & Cho, 2015; Pu et al., 2016). Moreover, the VAE framework has inspired the development of extensions targeting different aspects of feature representations and applications. Examples include importance-weighted AEs (Burda et al., 2016), the combination of a VAE with a generative adversarial network (Larsen et al., 2016), Wasserstein AEs (Tolstikhin et al., 2019), and Sinkhorn AEs (Patrini et al., 2020).

The inherently probabilistic nature of VAEs makes them potentially effective for the problem at hand. The standard VAE model is trained to learn a latent distribution for a single instance from the input data, where samples drawn from the latent distribution are decoded to data points close to the corresponding instance. If we consider each ensemble member separately, the VAE model would thus learn different latent distributions for different members from the same forecast case. Therefore, we need to adapt the VAE framework to jointly learn one latent distribution for all ensemble members and decode samples from the latent distribution to newly generated members that follow the same distribution as the inputs. To address this challenge, we propose an invariant VAE (iVAE) model, inspired by the permutation-invariant NN framework in the Deep Sets architecture (Zaheer et al., 2017). The encoder of our iVAE model follows such a permutation-invariant framework and is invariant to any permutation on the order of ensemble members. A schematic overview of our iVAE model is available in Figure 2.

The main difference between our iVAE model and a standard VAE lies in the inputs and the corresponding encoder architecture. Our iVAE model takes the full ensemble of 50 members as a single training instance and outputs a reconstructed ensemble of multiple members, thereby enabling probabilistic input and output in the empirical format. In contrast, a standard VAE processes only one ensemble member per training instance, analogous to the setup used in the PCA- and AE-based approaches, and its optimization still relies on the reconstruction error between deterministic input and output. The encoder of our iVAE is shared across all 50 ensemble members within a single input instance and is specifically designed to be invariant to the order of these interchangeable members. The shared encoder comprises two separate encoder parts, which we denote by e_1 and e_2; see Figure 2. For a given 50-member ensemble $\boldsymbol{X} = \{\boldsymbol{X}_m\}_{m=1}^{50}$, the first encoder is applied to each ensemble member forecast field $\boldsymbol{X}_m$ iteratively to obtain intermediate representations $\boldsymbol{y} = \{\boldsymbol{y}_m\}_{m=1}^{50}$, that is,

$$\boldsymbol{y}_m = e_1(\boldsymbol{X}_m), \qquad \text{for } m \in \{1, \ldots, 50\}.$$

These intermediate representations are interchangeable since their original input fields are assumed to follow the same distribution. Next, we average the 50 intermediate representations to summarize key features learned from all ensemble members, that is,

$$\bar{\boldsymbol{y}} = \frac{1}{50}\sum_{m=1}^{50} \boldsymbol{y}_m,$$

which ensures that the shared encoder is permutation-invariant. Note that other pooling operations such as maximum or minimum pooling could also be applied, though we did not observe improvements in our initial experiments. Subsequently, the second encoder is applied after this average pooling step. Similar to standard VAEs, a probabilistic encoder e_2 with parameters ϕ is applied to approximate the posterior distribution $p(z|\boldsymbol{X})$ in the latent space using a parameterized distribution $q_\phi(z|\boldsymbol{X}) = \mathcal{N}(z;\boldsymbol{\mu},\mathrm{diag}(\boldsymbol{\sigma}^2))$. After applying the reparametrization trick, we obtain

$$\begin{aligned} \boldsymbol{\epsilon}_n &\sim \mathcal{N}(\mathbf{0},\mathbf{I}), \\ (\boldsymbol{\mu}, \log\boldsymbol{\sigma}) &= e_2\left(\frac{1}{50}\sum_{m=1}^{50} e_1(\boldsymbol{X}_m)\right), \\ z_n &= \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}_n. \end{aligned}$$

The latent distribution $\mathcal{N}(z;\boldsymbol{\mu},\mathrm{diag}(\boldsymbol{\sigma}^2))$ is thus the low-dimensional probabilistic representation of the ensemble forecast fields that we aimed for. The decoder d with parameters θ is then applied to the sample z_n from the latent distribution to generate reconstructed forecast field $\tilde{\boldsymbol{X}}_n$, parameterizing the likelihood $p_\theta(\boldsymbol{X}|z)$ in the data space. In contrast to standard VAEs, we decode an arbitrary number N of samples $z = \{z_n\}_{n=1}^{N}$ for each data point, producing an ensemble of reconstructed forecast fields as output.

The architecture of our iVAE is built on the AE model discussed above and employs fully connected dense layers. The shared encoder adds an average pooling layer to the encoder of the AE model, consisting of e_1 with one layer of size "6,561 − 4,096" and e_2 with two layers of sizes "4,096 − 4,096 − d_{latent}," while the decoder follows the same structure as the decoder of the AE with layers of sizes "d_{latent} − 4,096 − 6,561." The LeakyReLU activation function is again applied in the hidden layers, and the same AdamW optimizer and early stopping criterion are applied. Due to the significantly increased memory requirements for training the iVAE model, we employ minibatch training with a batch size of 64.

The training objective of a standard VAE is to maximize the evidence lower bound on the marginal likelihood of the data,

$$\mathcal{L}(\theta,\phi) = \log p_\theta(\boldsymbol{X}|z) - D_{\mathrm{KL}}\left(q_\phi(z|\boldsymbol{X}) \| p_\theta(z)\right),$$

which consists of a negative reconstruction error and a regularization term. The reconstruction error $-\log p_\theta(\boldsymbol{X}|z)$ measures how well the model reconstructs the input data, which is proportional to the MSE with the Gaussian assumption on the data distribution for deterministic input and output. The regularization term is the Kullback-Leibler divergence between the approximate posterior $q_\phi(z|\boldsymbol{X})$ from the encoder and the prior $p_\theta(z)$ of the latent code z, where standard multivariate Gaussian distributions are often used as the prior.

In our case, however, the probabilistic nature of both input and output of the iVAE necessitates a different notion of the reconstruction error used for model training. Our iVAE model takes an ensemble of forecast fields $\boldsymbol{X} = \{\boldsymbol{X}_m\}_{m=1}^{50}$ as input and generates an ensemble of reconstructed fields $\tilde{\boldsymbol{X}} = \{\tilde{\boldsymbol{X}}_n\}_{n=1}^{N}$ with size N as output. The number N is not necessarily equal to 50, and each input member, $\boldsymbol{X}_m$, does not match the corresponding output member $\tilde{\boldsymbol{X}}_n$ for $m = n$ due to the random sampling of latent distribution. Therefore, the MSE between $\boldsymbol{X}_m$ and $\tilde{\boldsymbol{X}}_n$ is not a suitable choice for estimating the reconstruction error. Given that both the input and output ensembles of the iVAE can be considered to be multivariate empirical probability distributions, notions of the distance between the two distributions yield a more appropriate choice. We thus incorporate two such metrics into the iVAE training objective. Specifically, we use the energy distance and the Sinkhorn distance, which will be introduced in

Section 4.1, for measuring different aspects of distances between multivariate probability distributions. The reconstruction error of our iVAE is defined as the weighted sum of the energy distance and the Sinkhorn distance, complemented by the Kullback-Leibler divergence as a regularization term. The three loss components exhibit significantly different scales, necessitating rescaling to ensure that their value ranges are comparable. By comparing the mean values of different loss components over the first 20 epochs of training, we applied the following adjustments: the KL divergence was divided by 10, the energy distance was multiplied by 2 for both weather variables, and the Sinkhorn distance was divided by 50 for temperature and by 500 for wind speed. The loss function of our iVAE model for temperature data thus is

$$\ell(\boldsymbol{X},\tilde{\boldsymbol{X}}) = \omega_1 \cdot 2D(\boldsymbol{X},\tilde{\boldsymbol{X}}) + \omega_2 \cdot \frac{1}{50}\mathrm{SD}(\boldsymbol{X},\tilde{\boldsymbol{X}}) + \omega_3 \cdot \frac{1}{10}D_{\mathrm{KL}}\left(q_\phi(z|\boldsymbol{X})\,\|\, p_\theta(z)\right), \tag{1}$$

where $D(\cdot)$ represents the energy distance and $\mathrm{SD}(\cdot)$ denotes the Sinkhorn distance. This choice of loss function reflects a key conceptual difference from the PCA- and AE-based approaches, as the iVAE model treats all interchangeable ensemble members jointly as an empirical distribution in both the inputs and outputs and directly learns a distributional representation in the latent space. In our preliminary experiments, we observed that assigning a high weight to the KL divergence component in the loss function restricts the information flow through the bottleneck of the network, which results in outputs that fail to preserve the general spatial patterns of the input forecast fields. To mitigate this issue and alleviate posterior collapse, we, like many other studies, heuristically selected a small weight $\omega_3 = 0.01$ for the KL divergence component. For a more comprehensive understanding of the posterior collapse problem, we refer to Lucas et al. (2019). Regarding the two components used to measure reconstruction error, we assign equal weights of $\omega_1 = \omega_2 = 0.5$. Additional analyses on the sensitivity of the evaluation results to alternative weighting choices are provided in Supporting Information S1 (Chen et al., 2025). A more detailed investigation of factors such as ensemble spread, the representation of extremes, and the energy spectrum of perturbations under different weighting choices offers a natural extension of the current study. Beyond fixed weights, approaches allowing for dynamically adapted weighting schemes have been proposed (Clark Di Leoni et al., 2023; Foldes et al., 2024; Groenendijk et al., 2021; Heydari et al., 2019), which may offer further improvements and represent an interesting direction for future work.

4. Results

In the following sections, we first briefly introduce the evaluation methods tailored for our specific problem. Then we present and discuss the corresponding results for the dimensionality reduction methods introduced above. Finally, we analyze the encoded low-dimensional representations from different methods in an exemplary use case.

4.1. Evaluation Methods

Choosing appropriate evaluation methods for our specific setting presents a challenge. Two primary perspectives guide our evaluation: assessing the accuracy of the reconstructed ensemble forecast fields in comparison to the original ensemble fields and analyzing the information content of the encoded low-dimensional representations. Evaluating the latter is particularly challenging as there naturally is no ground truth information for the representations, and the suitability will strongly depend on the application use case. Therefore, our main focus is on discrepancy measures between the reconstructed output and the original input ensemble fields. The discrepancy could be evaluated in terms of independent pixel-wise errors at each grid point, and joint, whole-image evaluation, where the entire forecast field is considered at once. Several evaluation metrics are available for both settings and will be introduced in the following.

All three dimensionality reduction approaches encode a low-dimensional Gaussian distribution for representing an ensemble of forecast fields. We draw 50 samples from the encoded distribution and decode them into reconstructed forecast fields to enable a fair comparison when assessing the discrepancy with the raw 50-member ensemble. As discussed above, the reconstructed ensemble members do not necessarily match the individual raw ensemble members, which prohibits measuring the pairwise differences directly. As an alternative, we compare the mean and standard deviation of all ensemble members between the raw and reconstructed fields at each grid point, providing insight into how well the model captures general characteristics of the input ensemble. Given an ensemble of spatial forecast fields $\boldsymbol{X} = \{\boldsymbol{X}_m\}_{m=1}^{50}$ and an ensemble of reconstructed fields $\tilde{\boldsymbol{X}} = \{\tilde{\boldsymbol{X}}_n\}_{n=1}^{N}$ with $N = 50$, we compute the absolute difference of ensemble means and the standard deviation difference at each grid point (i,j),

$$e_{(i,j)} = \left| \frac{1}{50} \sum_{m=1}^{50} \boldsymbol{X}_{m,(i,j)} - \frac{1}{N} \sum_{n=1}^{N} \tilde{\boldsymbol{X}}_{n,(i,j)} \right|, \quad \Delta(\sigma)_{(i,j)} = \sigma\big(\boldsymbol{X}_{(i,j)}\big) - \sigma\big(\tilde{\boldsymbol{X}}_{(i,j)}\big),$$

where $i,j \in \{1,\ldots,81\}$ and $\sigma(\cdot)$ denote the standard deviation of the respective ensemble.

We further consider probabilistic measures to quantify the discrepancy between two distributions of the ensembles used as inputs and obtained as outputs. The evaluation of ensemble forecast fields could be executed pixel-wise, considering the (one-dimensional) univariate distribution at each grid point, or for the whole image, treating all grid points together as a high-dimensional multivariate distribution. Measuring the divergence between two distributions for evaluating climate models in the univariate setting has been studied by Thorarinsdottir et al. (2013), and here, we consider the distance measures for both univariate and multivariate settings, utilizing the energy distance and optimal transportation distances. The multivariate measures are also integrated into the reconstruction error component of the loss function for training our iVAE models, as discussed earlier in Section 3.3.

The energy distance introduced by Székely and Rizzo (2013) is a metric that measures the distance between two probability distributions. Following our notations in Section 3.1, consider an ensemble of forecast fields $\boldsymbol{X} = \{\boldsymbol{X}_m\}_{m=1}^{50}$ and reconstructed fields $\tilde{\boldsymbol{X}} = \{\tilde{\boldsymbol{X}}_n\}_{n=1}^{N}$ in $\mathbb{R}^{d_{\text{data}}}$ with $N = 50$, with the assumption that the ensemble members follow an identical distribution, that is, $\boldsymbol{X}_m \sim \mathcal{P}$ for $m \in \{1,\ldots,50\}$ and $\tilde{\boldsymbol{X}}_n \sim \tilde{\mathcal{P}}$ for $n \in \{1,\ldots,N\}$. The squared energy distance between $\mathcal{P}$ and $\tilde{\mathcal{P}}$ can be estimated in terms of expected pairwise distances between the two ensembles of samples,

$$D^2\big(\boldsymbol{X},\tilde{\boldsymbol{X}}\big) = \frac{2}{50N} \sum_{m=1}^{50} \sum_{n=1}^{N} \big\| \boldsymbol{X}_m - \tilde{\boldsymbol{X}}_n \big\| - \frac{1}{50^2} \sum_{m_1=1}^{50} \sum_{m_2=1}^{50} \| \boldsymbol{X}_{m_1} - \boldsymbol{X}_{m_2} \| - \frac{1}{N^2} \sum_{n_1=1}^{N} \sum_{n_2=1}^{N} \big\| \tilde{\boldsymbol{X}}_{n_1} - \tilde{\boldsymbol{X}}_{n_2} \big\|,$$

where $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^{d_{\text{data}}}$. The energy distance $D\big(\boldsymbol{X},\tilde{\boldsymbol{X}}\big)$ is negatively oriented and is zero if and only if the two empirical distributions with samples $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$ coincide. In the univariate setting of pixel-wise evaluation, the squared energy distance at each grid point (i,j) is thus

$$\begin{aligned} D^2_{(i,j)} = \frac{2}{50N} \sum_{m=1}^{50} \sum_{n=1}^{N} \big| \boldsymbol{X}_{m,(i,j)} - \tilde{\boldsymbol{X}}_{n,(i,j)} \big| &- \frac{1}{50^2} \sum_{m_1=1}^{50} \sum_{m_2=1}^{50} \big| \boldsymbol{X}_{m_1,(i,j)} - \boldsymbol{X}_{m_2,(i,j)} \big| \\ &- \frac{1}{N^2} \sum_{n_1=1}^{N} \sum_{n_2=1}^{N} \big| \tilde{\boldsymbol{X}}_{n_1,(i,j)} - \tilde{\boldsymbol{X}}_{n_2,(i,j)} \big|. \end{aligned}$$

The univariate squared energy distance is closely related to the Cramér distance (Rizzo & Székely, 2016), which is also known as the integrated quadratic distance (Thorarinsdottir et al., 2013) for evaluating probabilistic forecasts from climate models. The computation of univariate energy distance $D_{(i,j)}$ follows existing Python implementations from the `scikit-learn` library (Pedregosa et al., 2011), while the multivariate energy distance $D\big(\boldsymbol{X},\tilde{\boldsymbol{X}}\big)$ is implemented by custom code.

The optimal transportation distances, also known as the p-Wasserstein distances (Kantorovich, 1960), are another type of metric that measure distances between two probability distributions. The general optimal mass transport problem aims to find the optimal strategy to transport probability mass from one probability measure into another while minimizing the transportation cost, where the p-Wasserstein distance is the minimum total cost with the cost function $c(x,y) = |x - y|^p$. Optimal transportation distances have found broad applications in ML in recent years (Arjovsky et al., 2017; Courty et al., 2016; Frogner et al., 2015). For an overview of the theory and methodology of optimal transportation distances and their applications, we refer to Kolouri et al. (2017). In our univariate setting of pixel-wise evaluation, the empirical format of 1-Wasserstein distance is considered. The

1-Wasserstein distance between an input ensemble of forecast fields $\boldsymbol{X}$ and an output ensemble of reconstructed fields $\tilde{\boldsymbol{X}}$ at each grid point (i,j) is estimated based on order statistics,

$$W_{1,(i,j)} = \frac{1}{50}\sum_{k=1}^{50}\left|\boldsymbol{X}_{(k),(i,j)} - \tilde{\boldsymbol{X}}_{(k),(i,j)}\right|,$$

where the subscript $\cdot_{(k)}$ denotes the k-th order of values. We follow existing Python implementations from the `scikit-learn` library (Pedregosa et al., 2011) for the computation of univariate 1-Wasserstein distance $W_{1,(i,j)}$. In the multivariate setting of whole-image evaluation, estimating the Wasserstein distance between two high-dimensional distributions requires substantial computational cost and is thus not suitable for both evaluation tasks and particularly the integration into the iVAE training loss. The Sinkhorn distance proposed by Cuturi (2013) provides a computationally efficient approximation of the Wasserstein distance by leveraging entropic regularizations. We thus utilize the Sinkhorn distance as an alternative and follow the Sinkhorn algorithm proposed in Eisenberger et al. (2022) for a memory-efficient estimation. The MSE is used as the cost function, that is, $c(x,y) = \|x - y\|^p$ with $p = 2$. Thereby, the estimated Sinkhorn distance $\mathrm{SD}\big(\boldsymbol{X},\tilde{\boldsymbol{X}}\big)$ approximates the 2-Wasserstein distance between $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$,

$$W_2 = \inf_{\pi}\left(\sum_{k=1}^{50}\left\|\boldsymbol{X}_k - \tilde{\boldsymbol{X}}_{\pi(k)}\right\|^2\right)^{1/2},$$

where π denotes all possible permutations.

To compare different methods based on the same distance measure with respect to a benchmark, we further compute the associated skill score for analyzing the relative performances. Among our three dimensionality reduction approaches, the PCA-based approach is naturally taken as the reference baseline method. For each day in the test set, we first compute either the mean distance over all grid points for the pixel-wise metric or directly take the distance measure for the entire grid as the score S_a of a certain approach. The skill score SS_a is then calculated via

$$SS_a = \frac{S_{\mathrm{ref}} - S_a}{S_{\mathrm{ref}} - S_{\mathrm{opt}}},$$

where S_{ref} is the corresponding score of the reference PCA-based approach, and $S_{\mathrm{opt}} = 0$ represents the score of an optimal method, that is, an ideal model that replicates its input ensemble members perfectly. Skill scores are positively oriented and have an upper bound of 1, where a positive value indicates better performance than the benchmark, and a value of 0 indicates no improvement over the benchmark.

For brevity, we present only the results for the energy distances in the main text, whereas the results concerning the optimal transport distances are deferred to Supporting Information S1.

4.2. Reconstruction Accuracy

We compare the accuracy of the reconstructed ensemble forecast fields obtained by the three different approaches (i.e., PCA-based and AE-based two-step methods and the iVAE approach). In the interest of brevity, we here focus on 2-m temperature (`t2m`) and 10-m zonal wind speed (`u10`). The 10-m meridional wind speed and the geopotential height at 500 hPa were also investigated in initial experiments but showed generally similar results and are thus not discussed here.

Figure 3 shows examples of input forecast fields along with reconstructed ensemble members generated by the different approaches. To enable a clearer comparison across ensemble members, the ensemble mean has been subtracted from each member in both the raw and reconstructed data. As discussed above and in light of our aim of learning representations of the underlying probability distributions, the reconstructed ensemble members should not be expected to perfectly match the input even though the selected fields are from the same forecast day. For both target variables, all approaches are able to capture the general spatial structure in the raw fields despite

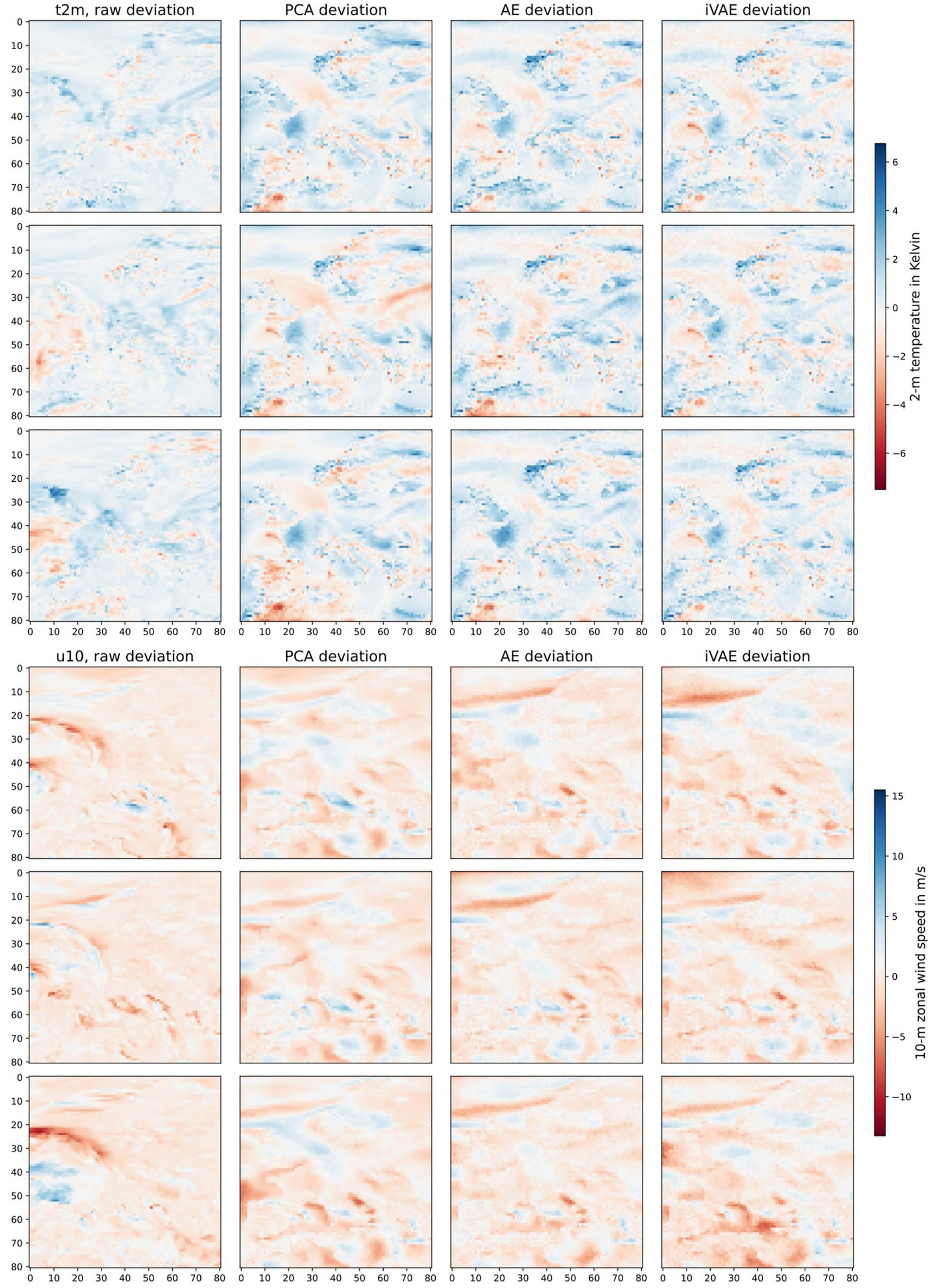

Figure 3. Exemplary raw forecast fields and reconstructed forecast fields of 2-m temperature (top) and the 10-m zonal wind speed (bottom) by different methods, substracting the ensemble mean with a latent dimension of 32. The rows correspond to three randomly picked ensemble members for the same forecast day.

reducing the dimensionality from 6,561 to 32, but neither of them is able to realistically replicate the localized fine patterns. The reconstruction quality and level of fine-scale details can be improved for higher-dimensional latent representation. Animated figures cycling through all ensemble members of reconstructed fields in comparison to

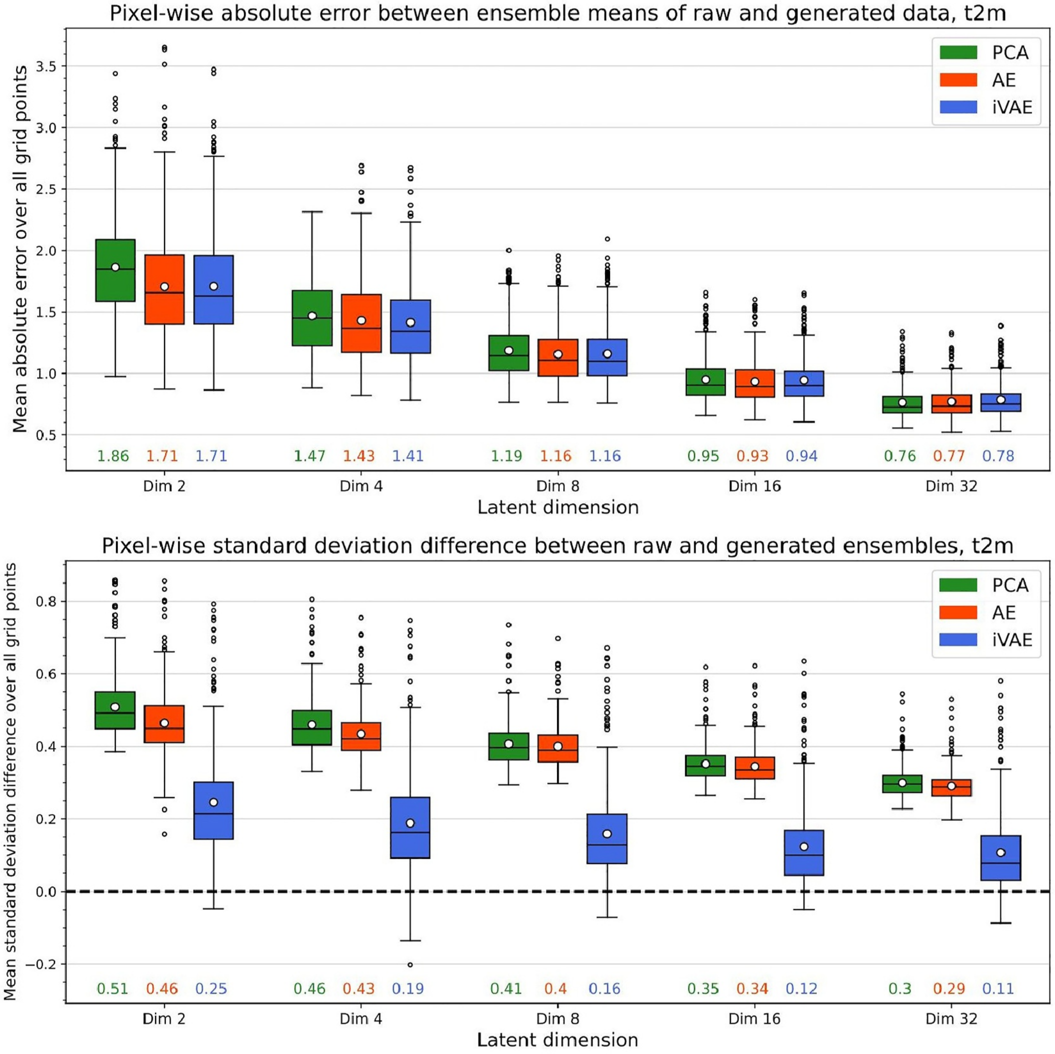

Figure 4. Boxplots of mean absolute differences between the mean values of input and reconstructed ensemble fields (top) and differences between the standard deviations of input and reconstructed ensemble fields (bottom) at each grid point. Boxes show performance variability over 366 days in the test set of different methods for 2-m temperature data, considering five different dimensionalities of the latent representation. The mean values of the (absolute) differences are indicated below each box. The differences between the standard deviations are computed such that negative values indicate a larger variability of the reconstructed ensemble compared to the input ensemble.

the corresponding ensemble of raw fields for the same forecast day, with or without substracting the ensemble mean, indicate that the iVAE method tends to reproduce larger but still realistic variability among different ensemble members compared to the other two methods. The animated figures are available in Supporting Information S1.

To assess the reconstruction quality in terms of general summary statistics of the input and output ensemble fields, Figure 4 shows pixel-wise absolute differences of the corresponding ensemble mean and standard deviation values for temperature data. For all methods, the differences between the summary statistics of the input and reconstructed ensemble fields decrease with increasing dimensionality of the latent representations. In terms of the deviations of the ensemble mean, the two NN-based methods show slightly better performance in lower-dimensional settings, while PCA shows minimally better performance for the largest dimensionality considered here. However, these differences are very minor, in particular when compared to the variability within the boxplots. More substantial differences can be observed for the standard deviation among the reconstructed ensemble members, where only the iVAE approach is able to produce variability among the members of a

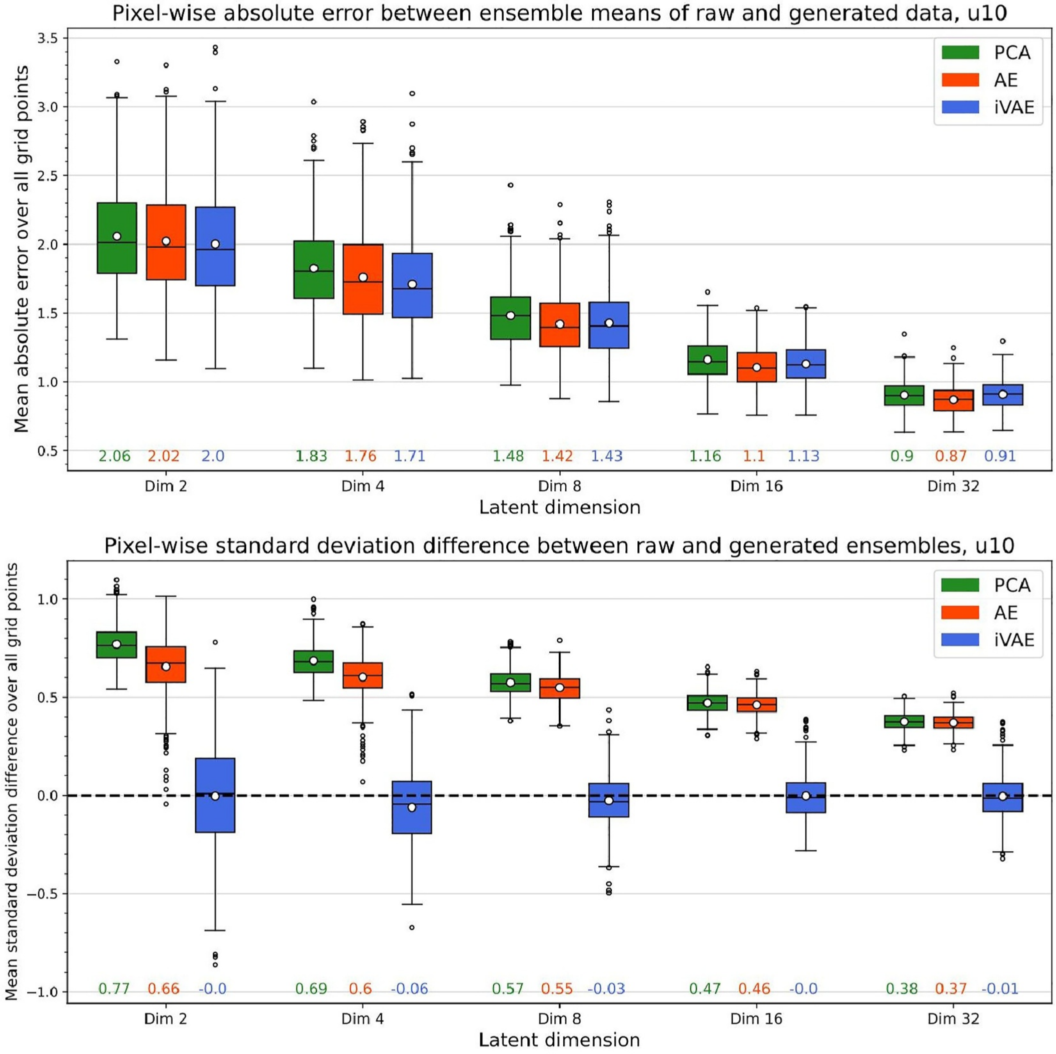

Figure 5. As Figure 4, but for 10-m zonal wind speed.

magnitude that matches the raw ensemble, whereas the reproduced ensembles from both two-step methods notably underestimate the variability of the input. Qualitatively similar results are obtained for the wind speed data, shown in Figure 5, where the better performance of the iVAE approach at correctly reproducing the variability across ensemble members is even more apparent.

For a more detailed assessment of the probabilistic reconstruction quality, Figure 6 shows skill scores based on energy distances between input and reconstructed temperature ensemble forecast fields for the AE-based and iVAE methods, with the PCA-based method as a baseline. Positive values indicate an improvement in terms of the energy distance over the reference method. For example, a skill score of 0.1 corresponds to a 10% lower energy distance than PCA. The iVAE method consistently outperforms the other two approaches across all latent dimensions and in both univariate and multivariate evaluation. These improvements are likely due to the variability of the reconstructed ensemble fields being close to the input ones as discussed in Figure 4 and potentially benefit from the inclusion of the multivariate energy distance as part of the loss function. The AE-based two-step approach generally outperforms the PCA-based approach, but the improvements are less pronounced than for the iVAE method, and poorer performance is observed for multivariate energy distances with a latent dimension greater than 8. Both NN-based approaches show less distinct improvements over the PCA-based method with increasing latent dimensions, possibly because PCA is sufficient to capture most of the variability information in the raw ensemble forecast fields when the dimension of representations is suitably large. Similar conclusions can

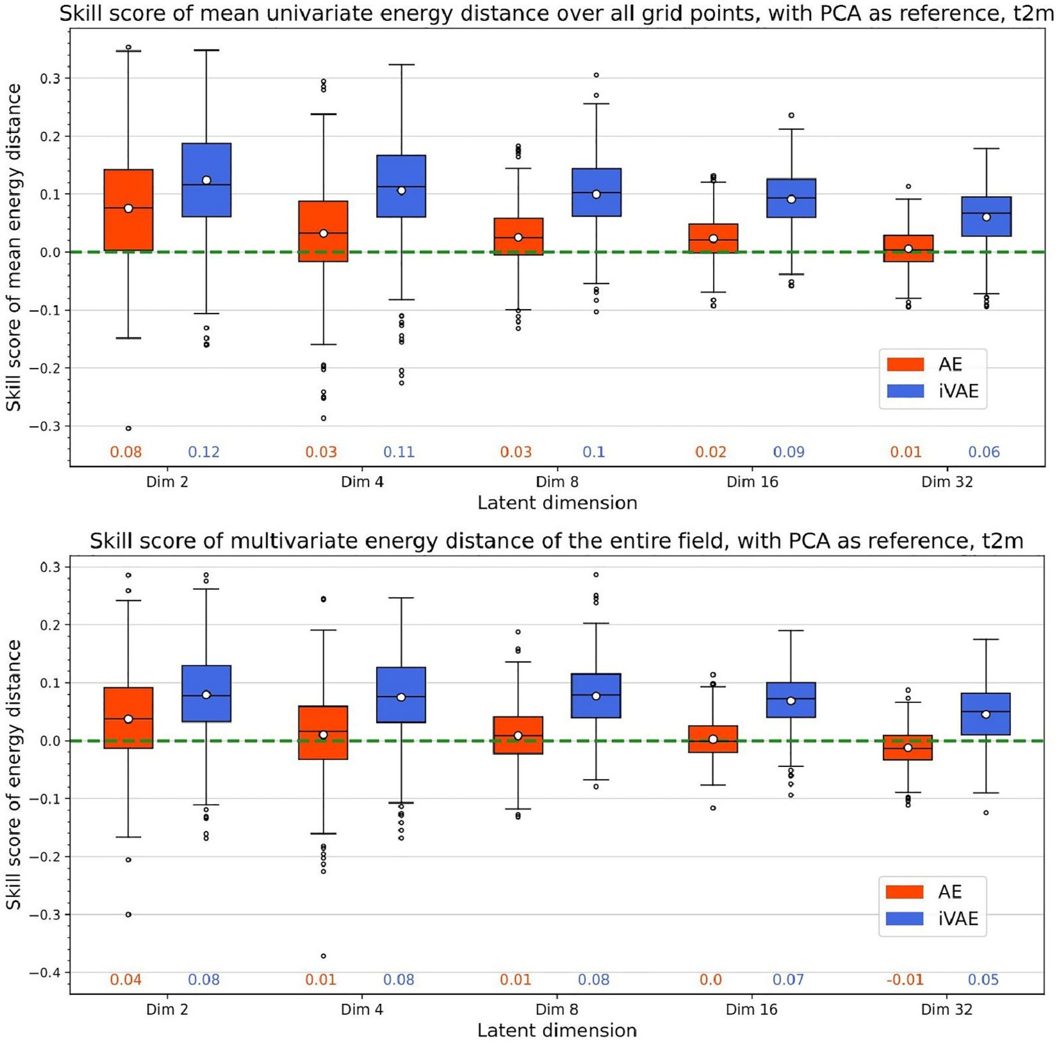

Figure 6. Boxplots of skill scores based on energy distances between the input and reconstructed ensemble fields over the 366 days in the test set for temperature data. The panels show mean univariate energy distances over all grid points (top) and multivariate energy distances computed for the entire fields (bottom). The principal component analysis-based approach shown in green dashed line is the reference method. The respective mean skill values are indicated below each box.

be drawn for the wind speed data; see Figure 7. The most notable difference to the results for temperature data is that the AE-based method consistently outperforms the PCA-based method here, and the largest improvements from the iVAE method occur at a latent dimension of 4.

4.3. Exploratory Analysis of the Encoded Representations

Here, we focus on another interesting question, which is whether the encoded representations in the low-dimensional latent space carry any relevant meteorological information or can offer additional insights about the data at hand. To that end, we focus on the temperature forecast data and try to detect seasonal patterns in the encoded representations. We restrict our attention to latent representations of dimension 2 to enable graphical visualization. Figure 8 shows scatterplots of the components of the mean vector of the encoded latent distributions for the different approaches.

The mean vectors of the latent representations from all three methods show clear patters corresponding to the seasonality of the 2-m temperature. In terms of the first coordinate on the *x*-axis, a clear separation into months typically associated with warmer or colder temperatures can be observed for all methods. The AE- and iVAE-

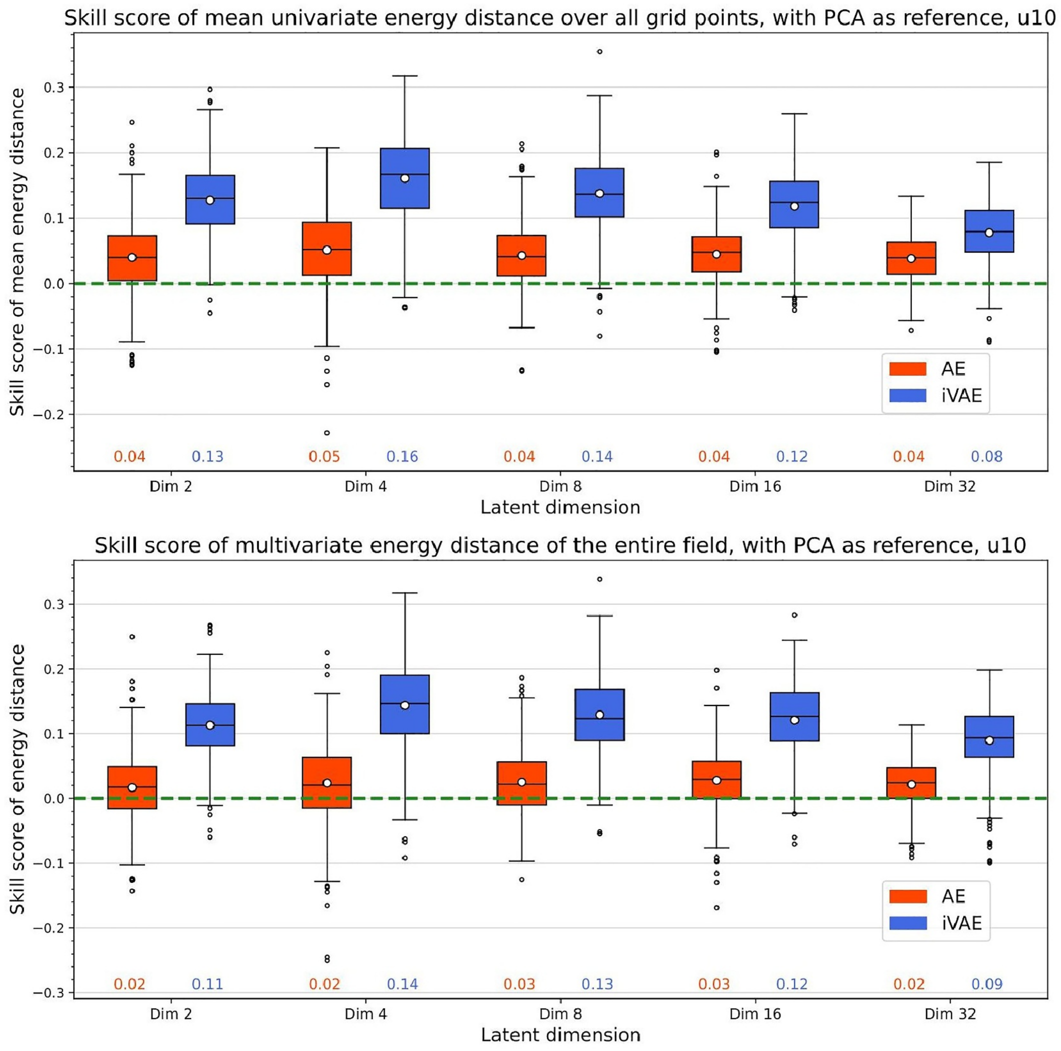

Figure 7. As Figure 6, but for 10-m zonal wind speed.

based representations further show some level of separation along the second coordinate on the *y*-axis, in particular between months in spring and autumn.

Similar scatterplots for the 10-m zonal wind speed are available in Supporting Information S1; however, there is a less clear seasonal pattern and a larger variability within individual months. We further explored the connection of the encoded representations of ensemble forecast fields of geopotential height at 500 hPa to quasi-stationary, recurrent, and persistent large-scale atmospheric circulation patterns, so-called weather regimes (Grams et al., 2017). While there are noticeable patterns and clusters which could be incorporated into postprocessing models (Mockert et al., 2024), the analysis is more involved and a more detailed investigation is left for future work. Some first results are available in Supporting Information S1. Further, exploring the information content of latent space representations using alternative techniques that have been proposed provides another promising starting point for future research; see, for example, Burgess et al. (2018), Locatello et al. (2019), and Rodríguez-Pérez and Bajorath (2020).

5. Discussion and Conclusions

We propose two types of approaches to learning probabilistic low-dimensional representations of ensemble forecast fields, which aim to treat them as interchangeable samples from an underlying high-dimensional probability distribution. The two-step framework based on traditional PCA or AE models treats each ensemble

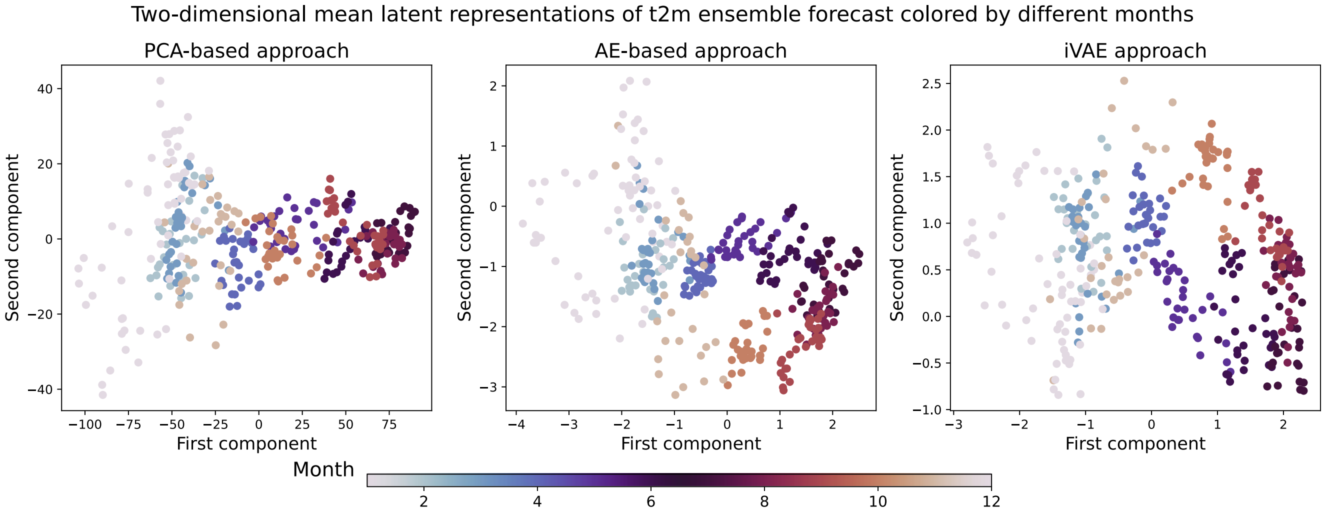

Figure 8. Scatterplots of the components of the mean vector of encoded two-dimensional representations of temperature for the three different dimensionality reduction methods. The points are colored according to the month of the corresponding forecast date.

member as an individual instance to learn deterministic latent representations and subsequently fits a multivariate probability distribution to the encoded representations of all ensemble members in the latent space. By contrast, the iVAE approach considers all ensemble members jointly as a single input and incorporates their interchangeable nature into the encoder design, thereby directly learning a probabilistic latent representation that captures the uncertainty present in the original ensemble. The proposed approaches offer two key conceptual advantages: first, they enable efficient dimensionality reduction of ensemble forecast fields within a probabilistic framework, where the encoded latent distributions encapsulate both structural and uncertainty information from the input ensemble; second, they allow for the computationally efficient generation of additional synthetic ensemble members beyond the size of the input ensemble, providing a straightforward way to expand the forecast ensemble and thereby improve uncertainty quantification. Systematic comparisons of PCA- and AE-based approaches and the iVAE in case studies on temperature and wind speed forecasts over Europe indicate that the two NN-based approaches show promising results both in terms of the quality of reconstructed forecast fields and the informativeness of encoded latent representations. While the results vary across all considered evaluation metrics, the iVAE model specifically performs best at preserving the variability information of the input ensemble forecasts. When compared to PCA, the NN-based methods generally show better performance for lower latent dimensions, whereas PCA yields equally good or even better reconstructions when the latent dimension is large. Overall, all three proposed methods require substantial memory capacity due to the large volume of the ensemble forecast data sets (approximately 9 GB for a single weather variable). Nevertheless, both the AE and iVAE models typically converge within 100 epochs. While the computational costs of training the iVAE are considerably higher than those of the AE- and PCA-based approaches, they remain manageable since the architectures are relatively shallow and consist solely of dense layers.

Despite the promising results, there are limitations to both types of approaches. All three methods assume a multivariate Gaussian distribution in the latent space, which might limit their applicability across different weather variables, specifically for variables such as precipitation, where the distribution should account for potential point masses at zero. A particular challenge in the specific setting of our dimensionality reduction problem is the evaluation of different approaches. In our experiments, we considered both deterministic and probabilistic metrics to assess the quality of reconstructed forecast fields. However, there is a general lack of suitable probabilistic evaluation tools that take into account structural aspects of the (ensemble of) forecast fields, compared to perception-based metrics proposed in the computer vision literature such as the widely used structural similarity index (Z. Wang et al., 2004). Spatial evaluation approaches proposed in the meteorological literature might offer useful starting points but are often heuristics-based, tailored to specific variables such as precipitation, and not straightforward to extend toward probabilistic settings; see, for example, Gilleland et al. (2010) and Dorninger et al. (2018) for overviews.

The proposed approaches provide several avenues for further generalization and analysis. Evidently, it would be of interest to investigate the scalability of the proposed methods toward larger grids and higher resolution forecast fields, as well as other variables. In particular, spatial forecasts of precipitation have been a focal point of research interest, for example, in spatial verification (Roberts & Lean, 2008). As noted above, the assumption of Gaussianity in the latent space would be a limitation, and adopting more flexible latent distributions, for example, through variational inference with non-Gaussian priors or normalizing flows, represents a promising direction for future work. Further, while we applied the dimensionality reduction methods separately to ensemble forecast fields of different variables, it would be interesting to apply them jointly to forecasts of multiple variables, and potentially over multiple forecast lead times. Previous results indicate that capturing small-scale variability, particularly in regions with complex terrain, remains challenging. Training models to reproduce the perturbations of each ensemble member relative to the ensemble mean may simplify this task and represents another potential avenue for further investigation. The normalization scheme could also be adapted to mitigate seasonal cycles, particularly for temperature, while forecast time information could be incorporated as additional predictors during model training. Furthermore, more advanced NN architectures, such as transformers, could be integrated as components of the AE-based or iVAE approaches to potentially improve performance. Given the recent developments in modern AI-based weather forecasting, the generation of AI-based ensemble forecasts (Bülte et al., 2025; Kochkov et al., 2024; Mahesh et al., 2024; Price et al., 2025; Zhong et al., 2024) might be another potential application of interest.

Finally, an important next step is to make progress toward integrating the proposed dimensionality reduction methods into downstream tasks. As discussed in the introduction, a key motivation for learning low-dimensional latent representations was to use those representations either as input data in, for example, hydrological or energy forecasting models, or to augment the input for NN-based postprocessing models (Lerch & Polsterer, 2022). Further examples of potential applications include (sub)seasonal weather prediction, where PCA-based representations of ensemble forecast fields have been used as inputs to ML models (Kiefer et al., 2023, 2024; Scheuerer et al., 2024), as well as analog forecasting, where compressed information from raw forecasts could be utilized for a more efficient generation of analogs (Grooms, 2021; Yang & Grooms, 2021). It is important to note that the quality requirements relevant to such applications differ from those in data reduction applications with a focus on data storage and management (see, e.g., Düben et al., 2019; Höhlein et al., 2022). While such approaches pursue accurate reconstruction as the central quality criterion, the achievable reconstruction quality may be decoupled from the information value of compressed representations in integrated modeling workflows. For example, in the specific context of incorporating encoded representations of spatial input fields into NN-based postprocessing models, one has to carefully balance the added value of the spatial information and the increased number of input predictors when the availability of training data is limited. In the setting of Lerch and Polsterer (2022), we did not find any improvements when using encoded representations of the full ensemble of forecast fields instead of the mean forecast only. One explanation in line with other findings from related work (Feik et al., 2024; Höhlein et al., 2024) might be that for postprocessing, there seems to often be little value of including full information from an ensemble beyond simple summary statistics.

Acknowledgments

The research leading to these results has been done within the Young Investigator Group "Artificial Intelligence for Probabilistic Weather Forecasting" funded by the Vector Stiftung. In addition, this project has received funding within project T4 "Development of a deep learning prototype for operational probabilistic wind gust forecasting" of the Transregional Collaborative Research Center SFB/TRR 165 "Waves to Weather" funded by the German Research Foundation (DFG). We thank Fabian Mockert and Christian Grams for providing weather regime data and helpful discussions, and thank Nina Horat, Benedikt Schulz, Uwe Ehret, and Tilmann Gneiting for helpful comments and discussions. We further thank three anonymous reviewers, whose constructive comments helped to improve an earlier version of this paper. This work is supported by the Helmholtz Association Initiative and Networking Fund on the HAICORE@KIT partition. The authors acknowledge support by the state of Baden-Württemberg through bwHPC. Open Access funding enabled and organized by Projekt DEAL.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

Python code with implementations of all methods in this study is available at https://doi.org/10.5281/zenodo.17170716. The data sets for the four variables considered are based on the underlying data from Rasp & Lerch (2018) and are available online (Chen and Lerch (2025a) for `t2m`, Chen and Lerch (2025c) for `u10`, Chen and Lerch (2025d) for `v10`, and Chen and Lerch (2025b) for `z500`).

References

Allen, S., Evans, G. R., Buchanan, P., & Kwasniok, F. (2021). Incorporating the north Atlantic oscillation into the post-processing of mogreps-g wind speed forecasts. *Quarterly Journal of the Royal Meteorological Society*, *147*(735), 1403–1418. https://doi.org/10.1002/qj.3983

An, J., & Cho, S. (2015). Variational autoencoder based anomaly detection using reconstruction probability. *Special lecture on IE*, *2*(1), 1–18.

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th international conference on machine learning* (Vol. 70, pp. 214–223). PMLR.

Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., & Baskurt, A. (2012). Spatio-temporal convolutional sparse auto-encoder for sequence classification. In *Bmvc* (pp. 1–12).

Bauer, P., Thorpe, A., & Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, *525*(7567), 47–55. https://doi.org/10.1038/nature14956

Bourlard, H., & Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics*, *59*(4–5), 291–294. https://doi.org/10.1007/bf00332918

Bülte, C., Horat, N., Quinting, J., & Lerch, S. (2025). Uncertainty quantification for data-driven weather models. In *Artificial intelligence for the earth systems*.

Burda, Y., Grosse, R., & Salakhutdinov, R. (2016). Importance weighted autoencoders. Retrieved from https://arxiv.org/abs/1509.00519

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., & Lerchner, A. (2018). Understanding disentangling in β-vae. Retrieved from https://arxiv.org/abs/1804.03599

Chapman, W. E., Monache, L. D., Alessandrini, S., Subramanian, A. C., Ralph, F. M., Xie, S.-P., et al. (2022). Probabilistic predictions from deterministic atmospheric river forecasts with deep learning. *Monthly Weather Review*, *150*(1), 215–234. https://doi.org/10.1175/mwr-d-21-0106.1

Chen, J., Janke, T., Steinke, F., & Lerch, S. (2024). Generative machine learning methods for multivariate ensemble postprocessing. *Annals of Applied Statistics*, *18*(1), 159–183. https://doi.org/10.1214/23-aoas1784

Chen, J., & Lerch, S. (2025a). Gridded dataset of daily 2-m temperature forecast from the ECMWF 50-member ensemble forecast. https://doi.org/10.6084/m9.figshare.28151213.v2

Chen, J., & Lerch, S. (2025b). Gridded dataset of daily geopotential height at 500 hPa forecast from the ECMWF 50-member ensemble forecast. https://doi.org/10.6084/m9.figshare.28151444.v1

Chen, J., & Lerch, S. (2025c). Gridded dataset of daily U component of 10-m wind speed forecast from the ECMWF 50-member ensemble forecast. https://doi.org/10.6084/m9.figshare.28151372.v1

Chen, J., & Lerch, S. (2025d). Gridded dataset of daily V component of 10-m wind speed forecast from the ECMWF 50-member ensemble forecast. https://doi.org/10.6084/m9.figshare.28151411.v1

Chen, J., Lerch, S., & Höhlein, K. (2025). Supporting information for "learning low-dimensional representations of ensemble forecast fields using autoencoder-based methods." (Available as supporting Information).

Clark Di Leoni, P., Agarwal, K., Zaki, T. A., Meneveau, C., & Katz, J. (2023). Reconstructing turbulent velocity and pressure fields from under-resolved noisy particle tracks using physics-informed neural networks. *Experiments in Fluids*, *64*(5), 95. https://doi.org/10.1007/s00348-023-03629-4

Courty, N., Flamary, R., Tuia, D., & Rakotomamonjy, A. (2016). Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(9), 1853–1865. https://doi.org/10.1109/tpami.2016.2615921

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 26). Curran Associates, Inc.

Dorninger, M., Gilleland, E., Casati, B., Mittermaier, M. P., Ebert, E. E., Brown, B. G., & Wilson, L. J. (2018). The setup of the mesovict project. *Bulletin of the American Meteorological Society*, *99*(9), 1887–1906. https://doi.org/10.1175/bams-d-17-0164.1

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. Retrieved from https://arxiv.org/abs/2010.11929

Düben, P. D., Leutbecher, M., & Bauer, P. (2019). New methods for data storage of model output from ensemble simulations. *Monthly Weather Review*, *147*(2), 677–689. https://doi.org/10.1175/mwr-d-18-0170.1

Eisenberger, M., Toker, A., Leal-Taixé, L., Bernard, F., & Cremers, D. (2022). A unified framework for implicit sinkhorn differentiation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 509–518).

Feik, M., Lerch, S., & Stühmer, J. (2024). Graph neural networks and spatial information learning for post-processing ensemble weather forecasts. In *International conference on machine learning 2024 - Machine learning for earth system modeling workshop*. Retrieved from https://arxiv.org/abs/2407.11050

Foldes, R., Camporeale, E., & Marino, R. (2024). Low-dimensional representation of intermittent geophysical turbulence with high-order statistics-informed neural networks (H-SiNN). *Physics of Fluids*, *36*(2), 026607. https://doi.org/10.1063/5.0179132

Frogner, C., Zhang, C., Mobahi, H., Araya-Polo, M., & Poggio, T. (2015). Learning with a Wasserstein loss. In *Proceedings of the 29th international conference on neural information processing systems* (Vol. 2, pp. 2053–2061). MIT Press.

Gilleland, E., Ahijevych, D. A., Brown, B. G., & Ebert, E. E. (2010). Verifying forecasts spatially. *Bulletin of the American Meteorological Society*, *91*(10), 1365–1376. https://doi.org/10.1175/2010bams2819.1

Gondara, L. (2016). Medical image denoising using convolutional denoising autoencoders. In *2016 IEEE 16th international conference on data mining workshops (ICDMW)* (pp. 241–246).

Grams, C. M., Beerli, R., Pfenninger, S., Staffell, I., & Wernli, H. (2017). Balancing Europe's wind-power output through spatial deployment informed by weather regimes. *Nature Climate Change*, *7*(8), 557–562. https://doi.org/10.1038/nclimate3338

Groenendijk, R., Karaoglu, S., Gevers, T., & Mensink, T. (2021). Multi-loss weighting with coefficient of variations. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 1469–1478).

Grönquist, P., Yao, C., Ben-Nun, T., Dryden, N., Dueben, P., Li, S., & Hoefler, T. (2021). Deep learning for post-processing ensemble weather forecasts. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *379*(2194), 20200092. https://doi.org/10.1098/rsta.2020.0092

Grooms, I. (2021). Analog ensemble data assimilation and a method for constructing analogs with variational autoencoders. *Quarterly Journal of the Royal Meteorological Society*, *147*(734), 139–149. https://doi.org/10.1002/qj.3910

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*.

Heydari, A. A., Thompson, C. A., & Mehmood, A. (2019). Softadapt: Techniques for adaptive loss weighting of neural networks with multi-part loss functions. Retrieved from https://arxiv.org/abs/1912.12355

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507. https://doi.org/10.1126/science.1127647

Hinton, G. E., & Zemel, R. (1993). Autoencoders, minimum description length and Helmholtz free energy. In J. Cowan, G. Tesauro, & J. Alspector (Eds.), *Advances in neural information processing systems* (Vol. 6). Morgan-Kaufmann.

Höhlein, K., Schulz, B., Westermann, R., & Lerch, S. (2024). Postprocessing of ensemble weather forecasts using permutation-invariant neural networks. *Artificial Intelligence for the Earth Systems*, *3*(1), e230070. https://doi.org/10.1175/aies-d-23-0070.1

Höhlein, K., Weiss, S., Necker, T., Weissmann, M., Miyoshi, T., & Westermann, R. (2022). Evaluation of volume representation networks for meteorological ensemble compression. In J. Bender, M. Botsch, & D. A. Keim (Eds.), *Vision, modeling, and visualization*. The Eurographics Association.

Horat, N., & Lerch, S. (2024). Deep learning for postprocessing global probabilistic forecasts on subseasonal time scales. *Monthly Weather Review*, *152*(3), 667–687. https://doi.org/10.1175/mwr-d-23-0150.1

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Kantorovich, L. V. (1960). Mathematical methods of organizing and planning production. *Management Science*, *6*(4), 366–422. https://doi.org/10.1287/mnsc.6.4.366

Kiefer, S. M., Lerch, S., Ludwig, P., & Pinto, J. G. (2023). Can machine learning models be a suitable tool for predicting central European cold winter weather on subseasonal to seasonal time scales? *Artificial intelligence for the Earth Systems*, *2*(4), e230020.

Kiefer, S. M., Lerch, S., Ludwig, P., & Pinto, J. G. (2024). Random forests' postprocessing capability of enhancing predictive skill on subseasonal time scales—A flow-dependent view on central European winter weather. *Artificial Intelligence for the Earth Systems*, *3*(4), e240014. https://doi.org/10.1175/aies-d-24-0014.1

Kingma, D. P. (2013). Auto-encoding variational bayes. Retrieved from https://arxiv.org/abs/1312.6114

Kingma, D. P., & Welling, M. (2019). An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, *12*(4), 307–392. https://doi.org/10.1561/2200000056

Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., et al. (2024). Neural general circulation models for weather and climate. *Nature*, *632*(8027), 1060–1066. https://doi.org/10.1038/s41586-024-07744-y

Kolouri, S., Park, S. R., Thorpe, M., Slepcev, D., & Rohde, G. K. (2017). Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, *34*(4), 43–59. https://doi.org/10.1109/msp.2017.2695801

Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, *37*(2), 233–243. https://doi.org/10.1002/aic.690370209

Kwok, J.-Y., & Tsang, I.-H. (2004). The pre-image problem in kernel methods. *IEEE Transactions on Neural Networks*, *15*(6), 1517–1525. https://doi.org/10.1109/tnn.2004.837781

Larsen, A. B. L., Sønderby, S. K., Larochelle, H., & Winther, O. (2016). Autoencoding beyond pixels using a learned similarity metric. In M. F. Balcan & K. Q. Weinberger (Eds.), *Proceedings of the 33rd international conference on machine learning* (Vol. 48, pp. 1558–1566). PMLR.

Lerch, S., & Polsterer, K. L. (2022). Convolutional autoencoders for spatially-informed ensemble post-processing. In *International conference on learning representations (ICLR) 2022 - AI for Earth and space science workshop*. Retrieved from https://arxiv.org/abs/2204.05102

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. (2018). Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, *18*(185), 1–52.

Liaw, R., Liang, E., Nishihara, R., Moritz, P., Gonzalez, J. E., & Stoica, I. (2018). Tune: A research platform for distributed model selection and training. Retrieved from https://arxiv.org/abs/1807.05118

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., & Bachem, O. (2019). Challenging common assumptions in the unsupervised learning of disentangled representations. In *International conference on machine learning* (pp. 4114–4124).

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. Retrieved from https://arxiv.org/abs/1711.05101

Lucas, J., Tucker, G., Grosse, R., & Norouzi, M. (2019). Understanding posterior collapse in generative latent variable models. Retrieved from https://openreview.net/forum?id=r1xaVLUYuE

Mahesh, A., Collins, W., Bonev, B., Brenowitz, N., Cohen, Y., Elms, J., et al. (2024). Huge ensembles part I: Design of ensemble weather forecasts using spherical fourier neural operators. Retrieved from https://arxiv.org/abs/2408.03100

Mika, S., Schölkopf, B., Smola, A., Müller, K.-R., Scholz, M., & Rätsch, G. (1998). Kernel PCA and de-noising in feature spaces. In M. Kearns, S. Solla, & D. Cohn (Eds.), *Advances in neural information processing systems* (Vol. 11). MIT Press.

Mockert, F., Grams, C. M., Lerch, S., Osman, M., & Quinting, J. (2024). Multivariate post-processing of probabilistic sub-seasonal weather regime forecasts. *Quarterly Journal of the Royal Meteorological Society*, *150*(765), 4771–4787. https://doi.org/10.1002/qj.4840

Patrini, G., van den Berg, R., Forré, P., Carioni, M., Bhargav, S., Welling, M., et al. (2020). Sinkhorn autoencoders. In R. P. Adams & V. Gogate (Eds.), *Proceedings of the 35th uncertainty in artificial intelligence conference* (Vol. 115, pp. 733–743). PMLR.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559–572. https://doi.org/10.1080/14786440109462720

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. R., El-Kadi, A., Masters, D., et al. (2025). Probabilistic weather forecasting with machine learning. *Nature*, *637*(8044), 84–90. https://doi.org/10.1038/s41586-024-08252-9

Pu, Y., Gan, Z., Henao, R., Yuan, X., Li, C., Stevens, A., & Carin, L. (2016). Variational autoencoder for deep learning of images, labels and captions. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 29). Curran Associates, Inc.

Rasp, S., & Lerch, S. (2018). Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, *146*(11), 3885–3900. https://doi.org/10.1175/mwr-d-18-0187.1

Rizzo, M. L., & Székely, G. J. (2016). Energy distance. *WIREs Computational Statistics*, *8*(1), 27–38. https://doi.org/10.1002/wics.1375

Roberts, N. M., & Lean, H. W. (2008). Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, *136*(1), 78–97. https://doi.org/10.1175/2007mwr2123.1

Rodríguez-Pérez, R., & Bajorath, J. (2020). Interpretation of machine learning models using Shapley values: Application to compound potency and multi-target activity predictions. *Journal of Computer-Aided Molecular Design*, *34*(10), 1013–1026. https://doi.org/10.1007/s10822-020-00314-0

Rodwell, M. J., Richardson, D. S., Parsons, D. B., & Wernli, H. (2018). Flow-dependent reliability: A path to more skillful ensemble forecasts. *Bulletin of the American Meteorological Society*, *99*(5), 1015–1026. https://doi.org/10.1175/bams-d-17-0027.1

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, *290*(5500), 2323–2326. https://doi.org/10.1126/science.290.5500.2323

Sakurada, M., & Yairi, T. (2014). Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *Proceedings of the MLSDA 2014 2nd workshop on machine learning for sensory data analysis* (pp. 4–11). Association for Computing Machinery.

Scheuerer, M., Heinrich-Mertsching, C., Bahaga, T. K., Gudoshava, M., & Thorarinsdottir, T. L. (2024). Applications of machine learning to predict seasonal precipitation for east Africa. Retrieved from https://arxiv.org/abs/2409.06238

Scheuerer, M., Switanek, M. B., Worsnop, R. P., & Hamill, T. M. (2020). Using artificial neural networks for generating probabilistic subseasonal precipitation forecasts over California. *Monthly Weather Review*, *148*(8), 3489–3506. https://doi.org/10.1175/mwr-d-20-0096.1

Schölkopf, B., Smola, A., & Müller, K.-R. (1997). Kernel principal component analysis. In W. Gerstner, A. Germond, M. Hasler, & J.-D. Nicoud (Eds.), *Artificial neural networks—ICANN'97* (pp. 583–588). Springer.

Schulz, B., & Lerch, S. (2022). Machine learning methods for postprocessing ensemble forecasts of wind gusts: A systematic comparison. *Monthly Weather Review*, *150*(1), 235–257. https://doi.org/10.1175/mwr-d-21-0150.1

Székely, G. J., & Rizzo, M. L. (2013). Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, *143*(8), 1249–1272. https://doi.org/10.1016/j.jspi.2013.03.018

Tenenbaum, J. B., de Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*(5500), 2319–2323. https://doi.org/10.1126/science.290.5500.2319

Thorarinsdottir, T. L., Gneiting, T., & Gissibl, N. (2013). Using proper divergence functions to evaluate climate models. *SIAM/ASA Journal on Uncertainty Quantification*, *1*(1), 522–534. https://doi.org/10.1137/130907550

Tolstikhin, I., Bousquet, O., Gelly, S., & Schoelkopf, B. (2019). Wasserstein auto-encoders. Retrieved from https://arxiv.org/abs/1711.01558

Vannitsem, S., Bremnes, J. B., Demaeyer, J., Evans, G. R., Flowerdew, J., Hemri, S., et al. (2021). Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big data world. *Bulletin of the American Meteorological Society*, *102*(3), E681–E699. https://doi.org/10.1175/bams-d-19-0308.1

Veldkamp, S., Whan, K., Dirksen, S., & Schmeits, M. (2021). Statistical postprocessing of wind speed forecasts using convolutional neural networks. *Monthly Weather Review*, *149*(4), 1141–1152. https://doi.org/10.1175/mwr-d-20-0219.1

Wang, W., Huang, Y., Wang, Y., & Wang, L. (2014). Generalized autoencoder: A neural network framework for dimensionality reduction. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) workshops*.

Wang, Y., Yao, H., & Zhao, S. (2016). Auto-encoder based dimensionality reduction. *Neurocomputing*, *184*, 232–242. https://doi.org/10.1016/j.neucom.2015.08.104

Wang, Z., Bovik, A., Sheikh, H., & Simoncelli, E. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612. https://doi.org/10.1109/tip.2003.819861

Yang, L. M., & Grooms, I. (2021). Machine learning techniques to construct patched analog ensembles for data assimilation. *Journal of Computational Physics*, *443*, 110532. https://doi.org/10.1016/j.jcp.2021.110532

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., & Smola, A. J. (2017). Deep sets. In I. Guyon, S. Gunn, M. Nikravesh, & L. A. Zadeh (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc.

Zhong, X., Chen, L., Li, H., Liu, J., Fan, X., Feng, J., et al. (2024). Fuxi-ens: A machine learning model for medium-range ensemble weather forecasting. Retrieved from https://arxiv.org/abs/2405.05925

Zhou, C., & Paffenroth, R. C. (2017). Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 665–674). Association for Computing Machinery.